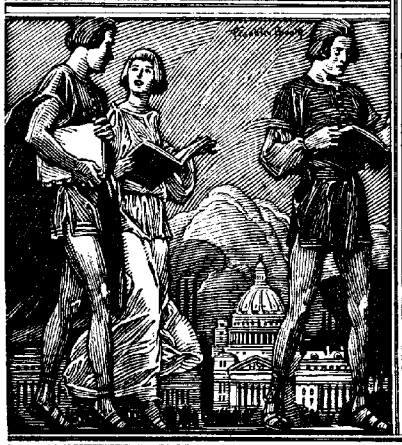


ELEMENTS OF STATISTICAL REASONING

BY

ALAN E. TRELOAR, PH.D.

ASSOCIATE PROFESSOR OF BIOSTATISTICS
UNIVERSITY OF MINNESOTA

NEW YORK
JOHN WILEY & SONS, INC.
LONDON: CHAPMAN & HALL, LIMITED
1939

PRINTED IN U. S. A.

PRESS OF
BRAUNWORTH & CO., INC.
BUILDERS OF BOOKS
BRIDGEPORT, CONN.

"It is always well to retain a clear geometric view of the facts when we are dealing with statistical problems, which abound with dangerous pitfalls, easily overlooked by the unwary, while they are cantering gaily along upon their arithmetic."

FRANCIS GALTON.

PREFACE

Much more than a trifle of truth is embodied in a simile drawn at the expense of many research workers who prop their argument with statistical proof. Likened to inebriated gentlemen affectionately clinging to lamp-posts, they are said to regard statistics as a source of support rather than as a center of illumination. One need not emphasize here the all too frequent appropriateness of the criticism. Rather, one is concerned with the frank admission of the critic that statistics do form a source of illumination for those whose wits are not dulled by number. Those whose intellectual activities find stimulus in the attainment of rigorous analysis turn toward statistical reasoning for illumination in the study of variation systems.

The steadily increasing acceptance of statistical methodology as a desirable tool in scientific analysis is reflected in the widening provision made in university curricula for the teaching of the subject. It has been the privilege of the author to be engaged in such instruction at the University of Minnesota for the past ten years. Built on foundations laid by Professor J. Arthur Harris, the exposition given in this book reflects the experience acquired during that period. The author's classes have been composed principally of graduate students with a great diversity of objectives. The urge to secure an understanding of the principles of statistical reasoning has, in these classes, made companions of entomologist and educator, of clinician and chemical engineer, of sociologist and mathematician, as much as of anatomist and geneticist.

An interest in the cultivation of logical analysis in science is all that has been asked as a prerequisite. Uncommon in these times is the student of biological phenomena who has been instilled with a keen appreciation of the analytical power in quantitative logic which a knowledge of mathematics may open to him; rare indeed is the one whose fortune it has been to acquire that power from formal courses in pure mathematics. Nevertheless, some aptitude for logical analysis in quantitative thinking is innate to most students who choose science as a prospective profession. A course in statistics may well be arranged to develop that aptitude through appeal to realities, whereas a prerequisite requirement of further training in theoretical mathematics oft-times conjures fears that effectively block further progress. Experience has amply shown that such

proficiency in mathematics as is finally demanded of the sound practical statistician may usually be developed quite well from but humble beginnings, and without recourse to current mathematical pedagogy.

To maintain the interest of the non-mathematically trained student, and to assist him in visualizing the nature of the problems under analysis, the method of geometric portrayal of situations by means of diagrams is freely employed. On that foundation, algebraic derivations of a simple nature may be developed without substantial loss of enthusiasm. Throughout the process the author has discerned from his classes that an appreciation of mathematics is inculcated, and some mastery of its reasoning processes is acquired.

Those who seek a compendium of statistical techniques, an array of formulas from which one or another may be culled to act as a mill for grinding fine flour from crude grain, merely by turning a handle, should not look further through these pages. What is written herein is intended for those who wish to reason carefully, not merely imitate. A sequence of such general concepts as seem to the author to be foundational is accordingly developed. From such a beginning it is believed that one may work forward to comprehension of the details of specialized procedures.

An original plan to include a critical analysis of small-sample techniques was finally laid aside in order that the discussion of elementary principles should not be too greatly restricted by the demands of space economy. The author believes that unfortunate consequences flow all too readily from an approach to statistical reasoning via the "small sample." It is not in the contemplation of meager sets of information that one most readily grasps general principles underlying the analysis of variation systems. If circumstances warrant it and time permits, the author will essay to complete the original plan in another volume.

The inclusion of an array of special exercises in this book has been deemed unnecessary. Only in its purely mathematical aspect does the statistical technique assume generality. Every set of factual data in science that may be assembled for statistical analysis presents questions which are intimately concerned with the validity of the data themselves. Misinterpretation all too readily flows from routine application of statistical procedures. Full understanding of the limitations of specific sets of data may usually be imparted only through lengthy discussion with those not familiar with the field. The author trusts that the reader will in general be able to provide himself with data in which he is interested and with the probable limitations of which he is reasonably familiar. Instructors may advisedly accumulate such material for practical assignments to meet the special interests of their classes. The reader who is restricted to the text as a source will, however, find opportunity to test

his grasp of procedure by using the various sets of data found there. It will be noted that, for the most part, these are studied with respect to one feature only, leaving ample room for other analyses.

It is very desirable that verification of all computations made in statistical work should be undertaken by the analyst himself. Frequently the reasonableness of the result of a calculation may be judged by careful inspection of the original data, coupled with some rough test by mental arithmetic. Such apparent concordance is, of course, not completely satisfactory in itself, but it will serve to indicate the presence of such gross arithmetical errors as all are subject to making in varying degree. Final verification through using a different arithmetic, such as a change of code scale or transfer to an alternative formula provides, is urged. The ease with which an error may be repeated in doing the same calculation over again before it has been completely forgotten is amazing. The statistical worker must learn to become self-reliant with respect to his arithmetic. Independent verification of results furnishes his most satisfactory attack on this problem.

The omission of a more complete bibliography of those writings which have contributed materially to the development of the statistical concepts dealt with is not accidental. For the most part, reference to them would be of very doubtful assistance to the reader to whom this discussion is addressed. In an elementary textbook of this nature one may in large measure and with equal justification present established concepts without historical review in much the same spirit as is characteristic of the elementary treatise on algebra.

The author is considerably indebted to others for such value as may pertain to this presentation. The memory of a brief association with Professor J. Arthur Harris is a constant impetus to aid the student of science in his advance along the path of clear thinking and more precise analysis. And the questions and observations of those students, whether ingenious or naïve, along with stimulating conferences with statistical colleagues and inspiring if somewhat difficult hours with the erudite writings of the great masters, have played their part to keep the path of teaching forever fresh. That others may have developed essentially the same type of exposition in their lectures is to be expected; indeed, the author is familiar with one case of striking parallelism. He has not, however, consciously borrowed the ideas of another without acknowledgment.

If the form of presentation proves as helpful to others as it appears to have been to the students with whom the author has worked, then the present undertaking of both author and publisher is justified. It is with a deep sense of gratitude that the many hours of most helpful discussions

with Dr. Borghild Gunstad are acknowledged by the author. Professor Lowell J. Reed, Dr. W. Edwards Deming, and Dr. Marian Wilder have contributed in like manner. The painstaking work of Miss Marjorie Moore in preparation of the diagrams, the curves for two of which were kindly supplied by Dr. Deming, is acknowledged with a sense of considerable indebtedness.

ALAN E. TRELOAR

UNIVERSITY OF MINNESOTA
February 9, 1939

CONTENTS

ELEMENTS OF STATISTICAL REASONING

CHAPTER 1

NUMERICAL DESCRIPTION

There have been many definitions of *science*, each one more or less colored by the particular interests of its promulgator and in some way failing to satisfy all. In a broad view, however, the essential objective of science is embodied in a brief statement which is quite familiar to those whose privilege it was to work with the pioneering American naturalist and biometrician, Professor J. Arthur Harris:

"The task of science is the description of the universe."

In offering this expression for consideration, one is not concerned here with the all-inclusive scope of science which is suggested, but with the method of science. It is the task of science to *describe*.

One may recognize immediately that the attainments of description motivated by scientific objectives vary widely; not all under its title necessarily warrants acceptance as science. Only description having the quality of undisputed acceptability, of winning universal assent to the truth of that set forth, achieves the ideal of scientific exposition. Acceptable formulations of the "laws of science," tracing fundamental relationships within the universe, arise only from description characterized by this quality.

SUBJECTIVE AND OBJECTIVE DESCRIPTION

The advance from obscure perception toward clearly defined understanding of the universe, physical and organic, has been made through increasing objectivity in meaningful description. Workers in some of the various divisions of science today have attained great skill in purely objective exposition, thereby approaching scientific ideals very closely. Others, though striving for the same ends, must still formulate their reasoning for the most part in a comparatively ill-defined language of purely verbal definitions, that is, in a highly subjective manner.

The varied degrees of accomplishment in objectivity of description characterizing the many fields of scientific inquiry may be considered a

consequence of the operation of two primary and interlocking factors. First, achievement in some sciences is fundamental to success in others; and, second, the complexity of phenomena in some fields must of necessity make satisfactory description very much more difficult than in others. In general, understanding of the elemental materials and forces of the universe and their relatively simpler interactions must pave the way to analysis of more complicated phenomena. In no small measure because of this the physical sciences have forged ahead over relatively smooth ground, whereas the basic biological sciences, and still further away the social sciences, fighting most complex entanglements of structural materials and multitudinous interacting forces, have found that at best they may advance but slowly in objective description.

Attention may be directed to the practical tasks involved in describing some well-known biological characters, such as *height* and *health* of human individuals. One essays to describe such characters because people differ with regard to them, and one wishes to make comparisons among people with respect to these variables. In a discussion with precision objectives it seems foolish at the present time to suggest describing human height merely by dividing all individuals into such ill-defined classes as are connoted by the words "tall" and "short." Nevertheless, some such move forms the first step in description. Subgroups are recognized within the whole, the simplest division being into only two classes. One might feel pleased to be able to describe, acceptably to others, the health of a large array of individuals by segregation merely into the two classes, "satisfactory" and "unsatisfactory." But of course one is in fact able to go much further, be more objective or "scientific," in describing height than in describing health. We may *measure* height in absolute units. The difference between the two situations is fundamentally one of degree of simplicity in the concepts of height and health, with attendant difference in degree of mastery of the problem of attaining precision in the description.

The evolutionary path of description for single characters is a passage from recognition of only two subdivisions in the range of variation, to recognition of more and more subdivisions giving 3, 4, 5, *et cetera*, more homogeneous groups qualitatively designated by suitable titles. Considering height so described, for instance, one may pause at 5 subdivisions for the moment and call them "very tall," "tall," "medium," "short," and "very short." Essaying to describe height thus in 100 individuals, one might study each one singly and finally manage to assign all to what seem to be the appropriate categories. In the absence of any measurement or of matching one against another, it would be a truly remarkable feat if there proved to be no overlapping of the groups with respect to the

statures embraced by them. Optical illusions might well cause some gross misplacements. Of far-reaching consequence, however, is the fact that, in the absence of measurement or direct comparison, errors of classification at group boundaries by any one individual must in general be expected to increase as the number of border lines between classes increases, that is, as the grouping becomes finer. Let many individuals essay independently to make such a classification of a single set of subjects, guided only by such descriptive terms as "very tall," *et cetera*, and the confusion when all results are assembled may be expected to become quite considerable.

There is no difficulty in imagining such a situation provided that the descriptive terms used are open only to subjective interpretation. Shift the variable under discussion from height to health, and one immediately faces a much more complex situation. Stature is an elementary variable portrayed by a single dimension, whereas health represents a composite of many unitary and interacting variables which for the most part are but dimly understood. What seems good health to a doctor is often believed to be poor health by a patient, and *vice versa*. That which comprises good health at one age may quite properly be regarded as poor health at another age. It is virtually impossible to reach general concordance of judgment in subdividing variation in the state of health into even very few subgroups. The use of descriptive terms open to varying subjective interpretations introduces of necessity an element of confusion into classification, whatever may be the character under study.

THE LANGUAGE OF NUMBER

Universal assent, or even a satisfactory approach to it, cannot be obtained without standardization of the terms in which description is made. The meanings of a large proportion of words and phrases used in qualitative description are susceptible of only temporary definition which at best may prevail rigidly within rather limited geographic bounds. Anti-vaccinationists today correctly quote Jenner himself as stating that smallpox vaccination may be accompanied by the development of "tumours" in the armpits! But the implication of the word "tumour" is quite different today from what it was in Jenner's time. Only in a relatively loose sense may subjectively interpretable words serve as a vehicle of scientific description. *Units of dimension*, on the other hand, have proved capable of universal and permanent standardization. They also have the property of unlimited divisibility without any loss of meaning, a quality quite foreign to words.

Description in terms of dimension allows one to pass directly from the

multitudinous verbal languages, with all their elements of subjective definition, to the single universal language of *number*, a language of unlimited scope and yet of absolutely fixed meaning. It is only natural that in their evolution all sciences have moved toward the greater and greater use of measurement and number as the vehicle of precise and objective definition. On this matter Quetelet [1] has remarked:

> The more advanced the sciences have become, the more they have tended to enter the domain of mathematics, which is a sort of center towards which they converge. We can judge of the perfection to which a science has come by the facility more or less great, with which it may be approached by calculation.

Number may arise in description with or without the use of measurement. Fundamental measurement may be defined as the counting of standard units and fractions thereof to fix a dimension. But entities or occurrences such as children per family, deaths among cases of typhoid fever, *et cetera*, which are perfectly definite in themselves, although indivisible, may be counted as precisely as units of measurement. Thus we have two clear-cut systems of numerical description. In one, the unit is fixed and divisible; in the other, it is definitely recognizable as an individual or an attribute of an individual, but it is indivisible. Both systems, however, provide a secure quantitative foundation for scientific inference.

VARIATION

The incentive to describe rests on the recognition of differences and the desire to distinguish clearly between variants. The taxonomist readily recognizes certain differences between two groups of plants or animals and uses those differences as a basis for the designation of two species. Then finer description shows that no two individuals in the one species are precisely alike. Is every plant or animal to be differentiated from every other one in taxonomy, or will it be adequate to recognize the existence and define the scope of variation within species?

The physical chemist, recognizing that elements differ, essays to find the atomic weight of each. In repeated determinations on any one he fails to secure completely concordant results. Are elements to be considered merely as classes of materials like biological species, within each of which variation may occur?

These two types of experience are found to be perfectly general throughout all the sciences. Variation seems universal; obvious and inescapable in some cases, it is quite illusive in others. With the attain-

[1] Helen M. Walker. Studies in the history of statistical method. Baltimore: Williams and Wilkins. 1929. *Vide* page 39.

ment of more precise levels in description through counting and measurement, and through the constant improvement in sensitivity of measuring devices, it is not surprising to find that qualities which may at first have seemed identical appear later to differ among themselves.

The two specific examples given above of variation within types serve to distinguish two general sources of variation. On the one hand, things may differ in fact; and, on the other, repeated measurements of a thing believed constant may lack perfect consistency. To designate these two classes of variation as "authentic" and "illusory" is not very practical. Every measurement is liable to error, and therefore "authentic" variation is never seen uncontaminated by the "illusory." In this latter connection it has pointedly been written:[2]

> No measurement of any real thing can ever be correct, for the simple reason that no instrument is capable of infinitely small displacements and no human eye can detect infinitesimal separations. Errors are therefore inevitable.

One may be content to recognize that variation in repeated measurements may be ascribed to two different sets of causes:

> (1) the inherent and environmental influences in response to which the character in question has attained its magnitude; and
>
> (2) the errors made by the observer and his measuring instruments collectively in attempting to determine the magnitude of the character.

Variation of one or both types must always be expected to underlie all observations. It is the objective of the statistical method to analyze such variation, and in view of that analysis to provide dependable bases for the formulation of certain types of inference.

Free communication of scientific thought and descriptive technique permits each worker in modern times, whatever his field may be, to profit quickly from the experience of other scientists. Precision methods of objective description developed in some one field now quickly influence accomplishment in others. As a result the demand for precise measurement, long established in the physical sciences, has grown apace in biological research. The quantitative data secured through counting or measurement in biology have, however, been frankly characterized by great variation in contrast to the relatively small variations observed in the "precise sciences." The data have called for new methods of analysis resting upon the secure foundation of mathematics, the logic of the

[2] R. W. M. Gibbs. The adjustment of errors in practical science. New York, Oxford University Press. 1929.

quantitative. These methods have been developed very largely by mathematicians attracted by the opportunities on biological frontiers. Designated as *biometric* or *statistical* methods of analysis, they have in recent years pervaded sciences wherein, before, only the most farsighted dreamed their presence could be tolerated, let alone be profitable.

PRECISION OF MEASUREMENT

The accuracy of mathematical deductions from data must inevitably be limited in some way by the precision and adequacy of the observations themselves. Emphasis must always be laid, therefore, on the necessity for the exercise of critical judgment in the collection and recording of data. Precision in measurement and care in classification materially enhance the value of observations.

It must not be inferred, however, that it is impossible to overdo these things. Overminuteness is mischievous. To measure adult stature to one-hundredth of an inch without determining the accuracy of the instrument used or the reliability of the observer's judgment, or without recording the physical state of the subject, is obviously foolish. Quite analogous overminute measurement in other connections is, however, not at all uncommon. False confidence in the validity of individual measures characterizes all too many pieces of work. It is effort well spent to determine accurately the trustworthiness of measurements and classifications as a special project before extensive recording is undertaken. Statistical experience in the analysis of errors of observation is an excellent teacher concerning misplaced confidence in measurement.

Ample room exists for more extensive exercise of critical judgment in determining the degree of refinement which is worth while in describing phenomena. Limits exist beyond which precision of measurement at the expense of number of observations is far too costly. One hundred measurements made to two significant figures may well be of much greater value to scientific analysis of a phenomenon than only ten measurements made to three or four significant figures. When the error of measurement is relatively small in comparison with the total range of variation of the measures, that error may well be quite unimportant. For the most part at least, it is far more consequential to define a variation system well by having many determinations, than it is to have the individual determinations made with great refinement in accuracy. That which is important to attainment of a broad scientific objective is too easily veiled by the personal satisfaction which great refinements in technique of execution of measurements may give.

THE DESCRIPTIVE VALUE OF MEASUREMENTS

As the complexity of phenomena increases, so the difficulty of determining measures which adequately portray each phenomenon increases. The task of determining the amount of iodine in a tissue is relatively simple when compared to that of measuring the intelligence of man. The latter variable is only in a limited way susceptible of quantitative evaluation with existing devices. Any measurements secured reflect only certain features of intelligence. One all too commonly encounters measurements which by terminology purport to describe rather fully the phenomenon to which they are applied when in fact the coverage is rather meager. One may venture the opinion that one of the most pronounced weaknesses in scientific inference, past and present, is overgeneralization by the would-be servants of science. Careful scrutiny of all measures to determine their limitations as descriptions of the phenomena being investigated is greatly needed among workers in the broad fields of biological science, particularly those engaged with problems in the social, psychological, and physiological behavior of human beings and their less fortunate proxies in experimental investigation.

THE SAMPLE AND THE SUPPLY

In gathering sets of measurements of a specific type, the scientist is motivated by a desire to learn something about the character measured, not alone in the individuals measured but in such individuals generally in the universe. Assuming that the measurement is appropriate to the character in question, and that it has been made with adequate accuracy, he has before him a description of the character in the individuals he has measured. Let it be assumed that the measurement is total milk yield from each of a number of cows of a specified breed during a fixed period following parturition. His interest in any use of those measurements does not pertain fundamentally to the individual animals tested, but to a much larger universe of measurements of which he believes his group to be representative. The particular group is just a *sample* from some indefinitely large *population, universe,* or *supply* of such measures, all of which he would prefer to have and use if there was no serious obstacle in the way of his doing so.[3]

Economy with respect to time if not finance must be expected to limit rather sharply the number of measurements that may be secured in the analysis of any problem. Accordingly, a sample is selected to form a

[3] The technical problem of the computational difficulty involved when enormous numbers of measurements must be handled is not pertinent here.

basis for broader interpretation. From the data comprising the sample the scientist wishes to draw conclusions concerning the supply of which that sample is representative. He wishes to make sound inferences with respect to a general body of data from reasonably precise knowledge of a relatively few individual measurements, a particular set. This is *inductive* reasoning, the inverse of deductive reasoning which passes from the general to the particular. New knowledge of the universe is always acquired through inductive inference from samples drawn from the universe, and statistical analysis forms a path along which such scientific inference may proceed with security.

As a basis for such inductive processes of thought, the samples must be representative of the supply in some specified manner. Accordance of the sample with such specifications is essential to the accuracy of the inductive process. The only generally adaptable specification of representativeness of a sample is that the individual measures comprising it should form perfectly random selections from the universe of reference. By perfectly random selection of a sample from a supply, one means, strictly speaking, that every individual in the supply has had equal opportunity of being drawn as a constituent member of the sample. All differences between samples thus drawn at random from the same supply are attributable solely to *errors of random sampling*. One of the prime functions of statistical methods is to determine the probability of any given difference between samples occurring because of such errors of sampling.

Random representativeness of samples is a prerequisite to the formal procedures of statistical analysis. If it should prove possible to select a sample that would be exactly alike in all characters to the universe of measurements from which it is drawn, then the statistical problems of induction would vanish and solutions concerning the populations could be given with certainty. Such a type of selective sampling from this universe is wholly impossible. It would require precisely that knowledge concerning the supply of measures which is unknown, determination of which forms the objective of the investigation.

In most practical research, the sample arises through certain circumstances rather than by any technique of random selection from a previously defined supply. In such cases, the supply which it may be considered to represent as a random sample must be defined by the investigator. This definition naturally arises from subjective judgment which is very likely to be colored by "wishful thinking." Limitations of the sample have not infrequently been overlooked in published investigations, and accordingly some conclusions have been drawn which are obviously spurious. Great care should be exercised to make sure that

statistical conclusions are drawn with reference only to the population of which the sample is truly representative in the sense just defined. The statistical method requires intelligent application to yield dependable results.

SYMBOLIC REPRESENTATION OF DATA

An invaluable contribution of mathematics to the sciences is its provision of a highly developed system of symbolic reasoning. The economy of thought and clarification of ideas which it makes possible constitute ample justification for its extensive use in the analytical reasoning of biology. In his introductory editorial to *Biometrika*, Karl Pearson wrote:

> . . . symbolic analysis widens our notions, it leads us at once to new points of view and it directly suggests fresh points for observation and novel directions for experimental research.

The truth of this statement is amply substantiated by the experience which all may readily acquire who blaze trails on the frontier of quantitative biology. Its power lies in simplification of statement, the portal to clear thinking.

Just what character is employed as a symbol for any entity is, of course, entirely optional. For convenience one usually chooses familiar symbols, such as the letters of alphabets and signs well enough known by usage to be generally known by name. There is not any necessity to retain a selected symbol for the same object of reference outside of the particular piece of analysis being conducted. Thus one may use the symbol x to represent the red blood cell count of healthy women students, or the protein content of samples of flour, or the calculated distance between two stars, or the number of days taken by a crop to reach the flowering stage, and so on. x may designate any character to which one wishes to refer repeatedly. Diversified interest will probably necessitate the use of several symbols in any one investigation, so that u, v, x, y, *et cetera*, may each represent a different entity. In such cases it is often helpful to employ symbols that are automatically suggestive in some way of the object designated.

In the following pages the symbols x, y, and so forth, designating variable quantities, are to be considered in a very general sense from the standpoint of research. As far as is convenient an effort will be made in general symbolism to follow the common mathematical practice of using the earlier letters of alphabets to designate constants and the later letters to designate variables. On occasions this convention proves hampering, and it may of course be dispensed with at will in favor of some more suitable alternative.

Let x, then, designate a chosen quantity which changes in magnitude from one individual to another, or from one determination to another on the same materials. It designates the variable character rather than any specific value of it. Such characters will be called *variables* herein. It is necessary to consider more than one value of x, "more than one cherry of the crop." Indeed, it is usually desirable to consider as many values of x as it is economically feasible to secure. The sequence of symbols,

$$x_1, x_2, x_3, \cdots, x_N,$$

may then represent in abbreviated form the series of N individual magnitudes comprising the sample. We shall herein call these individual values of x, the *variates*. It is to be noted particularly that the order of the subscripts does not necessarily imply order of magnitude. x_1 is just one of the N values of x forming the sample, x_2 is another magnitude, and so on. The series merely indicates that there are N variates in this sample of the variable x.

It may not be amiss to emphasize again just what the purpose is in measuring these N individuals. The scientific objective would be to find out what one can about them, not for their own sake but as a basis of generalization concerning the character x. These generalizations are to constitute a step in scientific description of the universe. To the owner of N cows the individual yields are a matter of considerable importance. He is interested in them as such for economic reasons. The scientist would be interested in the yields merely as an example of what such cattle in general may be expected to produce. Suppose these cows to be Jersey cows, and $y_1, y_2, y_3, \cdots, y_N$ to be milk yields for a herd of Guernsey cows. Is the milk yield of the Jersey breed higher than that of the Guernsey breed? In looking at the matter from the standpoint of science, one wishes to base a generalization on the observed values of x and y. One wishes to determine which is the superior breed in the characteristic of milk yield. The two series of records are but samples from two populations that are known to differ in many characters. It is desired to determine if possible whether that differentiation extends to the character in question.

SIZE OF SAMPLE

The dependability of any generalization which is made from impartial study of a sample drawn at random from a universe of discourse in which all magnitudes are not identical clearly must decrease as the number of individual measurements in the sample decreases, all other factors being equal. Paucity of information given by the basic data cannot be adjusted by mathematical corrections in the analysis of those data. The

trustworthiness of any generalization is a direct function of the number of cases upon which it is based. The desirability of avoiding the use of small samples when dealing with wide variation, except where it is not feasible to secure larger numbers of observations, should not be overlooked, for the hazards of interpretation intensify rapidly as N becomes very small. Detailed discussion of these hazards will be left for a later volume; the immediate concern is to stress the fundamental nature of the reasoning given in the first sentences of this paragraph. The student will do well to consider the point very carefully.

CHAPTER 2

THE LAW OF FREQUENCY DISTRIBUTION

Chaos is a state of lack of arrangement; that which appears to be chaotic may well be readily susceptible of orderly organization. Variation among physical and biological magnitudes appeared to scientists for a long time to be chaotic until study of methods of systematization of data showed that the quality of orderliness may be extended to the phenomenon of variation in all its forms. This revelation opened tremendously fertile fields for the expansion of quantitative description and the increase in precision of inference in the biological sciences.

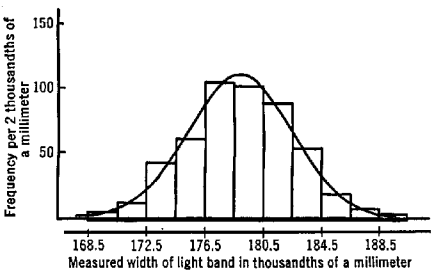
Orderliness in variation appears to have been first appreciated by the mathematical astronomer as a characteristic of data accumulated in the precise sciences. The gradation of errors made in the determination of spatial relationships of the heavenly bodies was observed by them to have a characteristic pattern coinciding with that of a mathematical function. Developed by the French mathematician, Laplace (1749–1827), and first applied to astronomical errors by the German mathematician and astronomer, Carl Friedrich Gauss (1777–1855), this function has become known variously as the "law of error," the "Laplace-Gaussian curve," or more recently as the "normal curve." It may well be called the principle cornerstone of the modern statistical edifice. Credit for discerning and primarily developing the application of this function to biological variations belongs to Adolphe Quetelet (1796–1874), the Belgian scientist distinguished for wide accomplishment growing from his tremendous breadth of interest. His dynamic sponsorship in Europe of this new "statistical" method of inquiry into biological and social phenomena was brilliantly followed in England by Francis Galton (1822–1911), whose enthusiasm was unbounded as he verified the principle of systematic variation in every biological variable for which he was able to accumulate adequate data.

Revelation of this principle of orderliness in biological variation formed the beginning of a new era in biological research, Galton writing with characteristic charm of the discovery in his epoch-making book "Natural Inheritance" (1889):

I know of scarcely anything so apt to impress the imagination as the wonderful form of cosmic order expressed by the "Law of Frequency of Error." The law would have been personified by the Greeks and deified, if they had known of it. It reigns with serenity and in complete self-effacement amidst the wildest confusion. The huger the mob, and the greater the apparent anarchy, the more perfect is its sway. . . . Whenever a large sample of chaotic elements are taken in hand and marshalled in the order of their magnitudes, an unsuspected and most beautiful form of regularity proves to have been latent all along.

FIGURE 1

FREQUENCY DISTRIBUTION OF 500 MEASUREMENTS OF THE WIDTH OF A SPECTRAL BAND OF LIGHT *



The "law of frequency of error" to which Galton referred, pertains to the change in frequency of occurrence of the successive magnitudes of a variable when they are placed in order. As one passes along the scale of measurement of a variable, from the smallest magnitude to the largest observed in any sufficiently extensive series, one finds an orderly change in the *frequency* with which each successive magnitude occurs. Most commonly the frequency of occurrence of each magnitude increases progressively with advance along the scale until a maximum is reached somewhere in the central region of the range and then an orderly falling away in frequency is observed until it finally disappears. A typical example of this characteristic in errors of measurement is provided in Fig. 1, wherein the frequency distribution of 500 determinations of the width of a spectral band of light is graphically presented. Among biological variables, anthropometric measures tend to follow fairly closely this symmetrical "law of error" function used first by Gauss.

* Data by courtesy of Professor Raymond T. Birge, University of California.

This is illustrated in Figs. 2*a* and 2*b*, where the same graduating curve used in Fig. 1 is applied to finger-length records secured from the Scotland Yard criminal identification files, and to scores made by a class of 449 students on an examination in "current affairs." In these three diagrams of frequency distribution the heights of the rectangles correspond to the frequencies observed within the designated ranges on the

FIGURE 2*a*

FREQUENCY DISTRIBUTION FOR LENGTH OF LEFT MIDDLE FINGER *

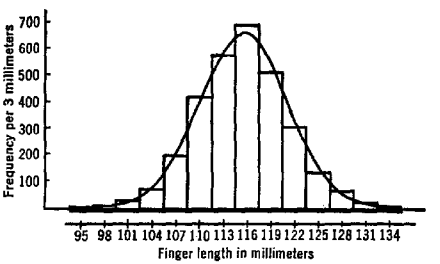
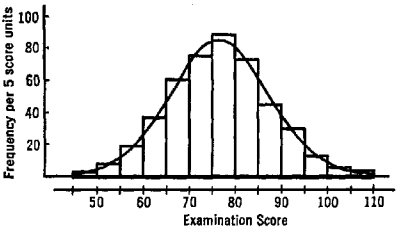


FIGURE 2*b*

FREQUENCY DISTRIBUTION FOR MARKS SECURED ON A "CURRENT AFFAIRS" EXAMINATION BY 449 STUDENTS †



scales of measurement given along the base. The curve in each case represents a mathematical smoothing of the polygon of rectangles. * It portrays the statistician's estimate of the frequency distribution in the assumedly infinitely large supply of such measures which conceivably might have been made in each case, and from which the samples may be considered as being assembled by random selection of individual measures.

* For basic data, see Table 3, page 19.

† Data by courtesy of the Committee on Educational Research, University of Minnesota.

Galton erred if he regarded the "law of frequency of error" as being descriptive of biological variations in general. Not all variables show this particular curve's specific characteristics in their frequency gradation. The type of mathematical function may change freely from one variable to another. However, all do show orderliness of frequency distribution in some related form. The *law of frequency distribution*, namely, that the frequency is a smooth mathematical function of the magnitude of a variable, seems truly universal in its sway. Let the variates be "marshalled in order" through seriation, and the form will unfold.

CONTINUOUS AND DISCRETE VARIABLES

The scales in terms of which the magnitudes of variables are numerically expressed are of two distinct general types, and biological variation may accordingly be classified into two categories on this basis. The egg production of the domestic fowl may be appropriately described in terms of the total weight of eggs laid, as well as in terms of the number of those eggs. The former is a continuous scale of measurement, theoretically capable of division into infinitesimal increments. The scale of weight proceeds continuously from any one specified unit to the next; there are no gaps or jumps. Any variable described in terms of such a scale may accordingly be designated as *continuous* in its variation.

The other scale is that of the integral numbers. It is not continuous but proceeds by discrete increments of one integer, or complete units. It is the basic scale of counting with which fractions are incompatible. Variables whose magnitudes are expressed on this scale of pure numbers are designated as *discrete* or discontinuous.

The general forms of distribution of frequency for both types of variable follow the same orderly patterns; the law of frequency distribution is independent of these scale characteristics. This will appear in that which follows.

SERIATION

The process of gathering together like magnitudes of a variable to form a frequency tabulation is known as *seriation, grouping,* or *classification.* It is a procedure of most simple character for any discrete variable, the scale of integers automatically defining for them a seriation scheme. Thus, using data quoted by Maynard,[1] the variation in number of children in completed families in New Zealand in 1916 is systematically portrayed by the adjacent seriation (Table 1) resulting from analysis of the original census data. Each possible value for number of

[1] G. D. Maynard. A study in human fertility. Biometrika, 14: 337–354. 1923

children per fertile marriage from one up to the maximum recorded (19) is recognized. Fractional values being impossible, this seriation involves no approximation; it is precisely definitive of all observed values.

TABLE 1

FREQUENCY DISTRIBUTION FOR NUMBER OF CHILDREN IN COMPLETED FAMILIES.* NEW ZEALAND, 1916

(Data from Maynard[1])

Number of children x	Number of families f
1	175
2	258
3	393
4	421
5	465
6	462
7	413
8	428
9	381
10	336
11	237
12	155
13	108
14	70
15	25
16	14
17	3
18	4
19	3
	4,351

In the case of any continuous variable, on the other hand, some unit of classification must be determined in advance and seriation made according to the arbitrary grouping so established. It is necessary to recognize clearly that the recorded magnitude of a variate on a continuous scale is only an approximation to its true value. The degree of such approximation is determined by the precision attained with the instruments employed in the measurement process. The recorded values of

* These data refer to fertile unions only, for mothers married at 20 years of age who, at the time of census, were between 45 and 65 years of age. The families may reasonably be considered to be completed, any corrections being of trifling consequence to the form of the distribution.

the variable appear to be discontinuous or proceed in "jumps," those jumps being equal to the smallest increment or graduation recognized in making the measurement. This increment therefore determines the smallest possible unit of classification of a series of variates on a continuous scale. It automatically imposes a classification on the truly continuous scale.

We may consider the head breadth measurements made on 1,000 Cambridge University students [2] as illustrative material. These measurements were made to the nearest tenth of an inch, and ranged from

TABLE 2

FREQUENCY DISTRIBUTION FOR HEAD BREADTH IN CAMBRIDGE MEN

(Data from Macdonell[2])

Head breadth in inches x	Frequency f
5.5	3
5.6	12
5.7	43
5.8	80
5.9	131
6.0	236
6.1	185
6.2	142
6.3	99
6.4	37
6.5	15
6.6	12
6.7	3
6.8	2
	1,000

a minimum of 5.5 inches to a maximum of 6.8 inches. The seriation imposed by the increment of measurement is given in Table 2. Since measurement was made to the *nearest* tenth of an inch, the 236 cases recorded as 6.0 inches, for instance, really fall over the range 5.95 to 6.05 inches. No attempt was made to differentiate them, all being recorded as the same magnitude.

More often than not, measuring instruments commonly used in the biological sciences are graduated and read to very small increments in

[2] W. R. Macdonell. On criminal anthropometry. Biometrika, 1: 177–227. 1902.

comparison with the total range of variation encountered. Under such conditions the increment of measurement provides a seriation scheme with a very large number, probably an altogether unwieldy number, of classes in which frequency is to be recorded. It then becomes advantageous to establish classes of broader range for seriation purposes, classes that have a range of some simple and preferably constant multiple of the increment of measurement. Selection of that multiple will be quite arbitrary, and may be made according to the purposes which the seriation is to serve. A case involving a small multiple may be used by way of illustration.

In the study just referred to, Macdonell [2] also gives data abstracted from the records of the Central Metric Office, New Scotland Yard, for length of left middle finger (measured to the nearest millimeter) in 3,000 prisoners. Seriations of these data are given in Table 3 in two classifications: (1) according to that imposed by the increment of measurement, and (2) in broader classes of three such intervals. Although the recorded measurements proceed by discrete steps of 1 mm., finger length is of course a continuous variable. A recorded value of 95 mm., when properly interpreted, means merely that the true value lies within the range of 94.5 to 95.5 mm. If the measuring instrument had been graduated only in increments of 3 mm., starting, say, at 92 mm., then the records would have followed the seriation given on the right-hand side of the table. Thus, grouping of data into broader classes than those imposed by the increment of measurement is simply the equivalent of having used a coarser scale of graduations in measurement.

It will be clear that measurement of any continuous variable involves errors of approximation, and seriation of such measurements into classes of broader range will have the effect of magnifying those errors. Their influence, however, may be reduced to an essentially negligible level by appropriate correction when the refinement of the investigation warrants it. These "corrections for grouping" are rarely of importance in biological research, particularly if there be, say 10 or more classes in the final seriation scheme.

Seriation plays a very important role of initial simplification in the statistical reduction of data. It permits of the replacement of N individual measurements by a very much smaller number of classes of measurements. In the absence of a calculating machine, seriation may be a decided help in computations for all but very small series of values. The number of classes which is desirable is a matter for individual judgment based on the nature of the problem in hand. It might be stated as a rough guide that from 8 to 15 classes are suitable for most practical purposes, with preferably an average frequency per class of 10 or more.

TABLE 3

FREQUENCY DISTRIBUTION FOR LENGTH OF LEFT MIDDLE FINGER IN CRIMINALS

(Data from Macdonell[2])

Recorded value in mm.	True range	Frequency	Class range	Class center x	Frequency f
94	93.5– 94.5	0			
5	94.5– 95.5	1	93.5– 96.5	95	1
6	95.5– 96.5	0			
7	96.5– 97.5	0			
8	97.5– 98.5	1	96.5– 99.5	98	4
9	98.5– 99.5	3			
100	99.5–100.5	7			
1	100.5–101.5	7	99.5–102.5	101	24
2	101.5–102.5	10			
3	102.5–103.5	17			
4	103.5–104.5	20	102.5–105.5	104	67
5	104.5–105.5	30			
6	105.5–106.5	44			
7	106.5–107.5	74	105.5–108.5	107	193
8	107.5–108.5	75			
9	108.5–109.5	102			
110	109.5–110.5	163	108.5–111.5	110	417
1	110.5–111.5	152			
2	111.5–112.5	183			
3	112.5–113.5	164	111.5–114.5	113	575
4	113.5–114.5	228			
5	114.5–115.5	233			
6	115.5–116.5	226	114.5–117.5	116	691
7	116.5–117.5	232			
8	117.5–118.5	184			
9	118.5–119.5	162	117.5–120.5	119	509
120	119.5–120.5	163			
1	120.5–121.5	126			
2	121.5–122.5	91	120.5–123.5	122	306
3	122.5–123.5	89			
4	123.5–124.5	44			
5	124.5–125.5	52	123.5–126.5	125	131
6	125.5–126.5	35			
7	126.5–127.5	31			
8	127.5–128.5	25	126.5–129.5	128	63
9	128.5–129.5	7			
130	129.5–130.5	8			
1	130.5–131.5	2	129.5–132.5	131	16
2	131.5–132.5	6			
3	132.5–133.5	2			
4	133.5–134.5	0	132.5–135.5	134	3
135	134.5–135.5	1			
		3,000			3,000

As will be shown shortly, seriation with broad grouping is often necessary for appropriate graphical representation of frequency distributions. When a calculating machine is available, it may not be advantageous to bother about seriation for purely computational purposes unless the number of measurements is of the order of 100 or more.

There remain three matters affecting seriation which certainly warrant some remark in this brief discussion.

(A) *Broad grouping of a discrete variable.* One may on occasion choose to seriate a discrete variable into classes of range greater than the unit increment of the number scale. One may consider the red blood cell count in man, by way of illustration. Usually made per ten-thousandth of a cubic millimeter of blood, these may vary over a range of several hundred cells between the smallest and largest count. Seriation by integers would obviously give too many groups. One may well use ranges of 20 or more blood cells to a class to bring the number of classes down to a reasonable figure.

(B) *Irregular grouping.* A constant range for all classes in the seriation scheme has very definite simplification value in computational procedure, but it is by no means necessary at all times. Indeed, it is common practice in social statistics to use irregular grouping for many variables. Table 4 gives a type of irregular classification commonly appearing in vital statistics, wherein the groupings may be expected to be dictated in part by homogeneity of interest within them from a disease point of view. In the statistical study of his data the student of vital statistics may readily free himself of the handicaps of irregular grouping by making comparison between groups in terms of "rates," to which some attention will later be given.

(C) *True class range.* In the seriation of continuous variables, the range of each class appears to be the interval from the lowest to the highest recorded measurement for each class. This is not the full range of the class on the continuous scale, however, for it would leave vacant gaps between the ending of one class and the beginning of the next. Since each recorded measurement actually refers to a range about a reference point, allowance must be made for the increment of measurement in establishing the *true* class range or class center.

Where measurement is made to the nearest graduation on the scale, the true class range is the apparent range extended at each end by one-half of the increment of measurement. Some workers, however, prefer to ignore the final fractional increments in measuring; they record the graduation below rather than the "nearest" graduation. In stating age it is a common practice for all to give the age *last* birthday rather than to determine the nearest birthday. This method of recording has

the obvious advantage of avoiding the exercise of subjective judgment as to which is the nearest graduation (or the effort of computing which is the nearest birthday) when the true value falls in the mid-region between graduations. On the other hand, when a graduation falls very close to the true value, precise measurement may be demanded to determine whether that graduation is really below or above the true value. When the terminal fractional increment is ignored in measurement, it must be

TABLE 4

DISTRIBUTION BY AGE OF THE WHITE POPULATION OF MINNESOTA AS RECORDED IN THE 15TH U. S. CENSUS, 1930

Age in years	Population	
	Native born	Foreign born
Under 1	43,397	10
1– 4	184,407	211
5– 9	252,492	1,419
10–14	249,364	2,013
15–19	233,032	4,670
20 24	203,628	8,781
25–29	177,780	13,747
30–34	169,946	17,985
35–44	291,640	70,680
45–54	176,938	89,811
55–64	102,277	81,459
65–74	50,241	66,376
75 and over	14,728	30,994
Unknown	809	138
Total	2,150,679	388,294

observed that the recorded value indicates that the true value lies between the recorded value and the next higher graduation on the scale of measurement. In such cases the true range of a class in a seriation scheme is the apparent range extended at the *upper* limit by one increment of measurement. In Table 4, for example, the true class ranges are 0–0.9̇, 1–4.9̇, 5–9.9̇, $\cdots$, and the class centers are 0.5, 3.0, 7.5, $\cdots$.

Errors of approximation balance one another when measurement is made to the nearest graduation, whereas those errors give a negative bias to all measurements made to the graduation below. A desire to record a value as near as feasible to the true value seems to express itself

naturally in general practice. In the absence of standardization concerning the disposition of fractional increments in practical measurement, the statistician must be alert to see that he ascribes the true class center and range to each frequency.

SERIATION PROCEDURE

The technique of seriation commonly employed for simple series of data recorded on sheets is to make in a suitable blank table a consecutive list of the seriation classes decided on, then enter each variate in its appropriate class by a short vertical tally stroke. The use of a cumulative cross stroke for every fifth tally in a class aids the final addition. The seriation of weight per bushel in 59 samples of wheat by this method is illustrated in Table 5. This tally method is likely to be quite tedious,

TABLE 5

ILLUSTRATING SERIATION OF A LIST OF 59 WEIGHTS PER BUSHEL OF WHEAT

Apparent class range	Tally	Class frequency f	Class center x
55.0–55.4	//	2	55.2
55.5–55.9	////	4	55.7
56.0–56.4	///	3	56.2
56.5–56.9	///	3	56.7
57.0–57.4	𝍸 𝍸 𝍸 /	16	57.2
57.5–57.9	𝍸 ////	9	57.7
58.0–58.4	𝍸 𝍸 /	11	58.2
58.5–58.9	𝍸 /	6	58.7
59.0–59.4	//	2	59.2
59.5–59.9	///	3	59.7
		59	

however, with large series of data. It also suffers the disadvantage of requiring a complete duplication of the work to secure a check on the accuracy of the seriation.

A much more satisfactory procedure is achieved by sorting cards having the raw data entered on them in a suitable form. This is particularly true when two or more variables are to be correlated. In such cases it is well worth while to design a suitable card first and transcribe all the data onto such cards. The N observations on the primary variable are entered one each on a separate card, and with these are also entered the associated values for any other variables to be considered in

the analysis. With several variables, cards may be specially printed with the variable designations if the project warrants it. The card reproduced in Fig. 3 is illustrative. It was prepared for a series of several thousand cases of obstetrical confinement, 37 variables in all being considered.

FIGURE 3

A Statistical Record Card with Printed Variable Designations, Prepared for an Obstetric Study

M. G. H. Case No.	30-2012			Series	1930
MOTHER: Age	20	F. Menst.	14	CHILD: Sex	F
Ht.	66	Interval	R. ~~Irr.~~	Lth.	50
Wt. A. P.	103	Int. days	5-6 weeks	Wt.	3475
Wt. P. P.	104.5	Duration	4-5	M. Age	288
Intersp.	26·5	Pains	B Ø X	Gravida.	2
Intercr.	28·5	Severity	—	Para.	2
Rt. Obl.	24·5			Pos.: In.	ODA
Lft. Obl.	25	L. Menst.	5-28-29	Pos.: Out.	ODA
Ext. Conj.	20	Qck.	4·5 m.	N. B. Heart	—
Diag. Conj.	12·5	Labor I	6:45	Fetal Heart	148
Obs. Conj.	—	II	0:44		
Intertub.	10	III	0:35	Am. x ?	
Post Sag.	8.5	Date Del.	3-12-30	Catholic	OWA

When only very few variables are to be studied, cards with printed titles are by no means necessary. A key card may be prepared to define the positions for entry of the data on blank cards, of the same size, as in the following specimens. Figure 4 shows the key card for a set of three chemical determinations made independently by two analysts on a series of flour samples. Figure 5 reproduces the record card for Sample 1895 of the series. There can be no confusion here as to the identification of the numbers so long as the key card is preserved.

It is a very simple matter to sort such cards into stacks on a flat surface according to some seriation scheme. Each stack may be checked separately with ease to see that its variates fall within the specified class range. The number of cards in each stack may then be counted and the frequency entered directly into its proper place in a prepared table. The cards may readily be preserved in seriation by cross-stacking if desired, being then ready for reseriation within each group for a second variable to be correlated with the first, and so on.

Tremendous strides have been made in the last decade or two in the development of mechanical recording and card-sorting systems. These depend primarily on careful planning of a standard sized card to be

FIGURE 4

A Specimen "Key Card" for Identification of Entries on Similar Cards Without Inscribed Variable Designations (Corresponding Record Card Below)

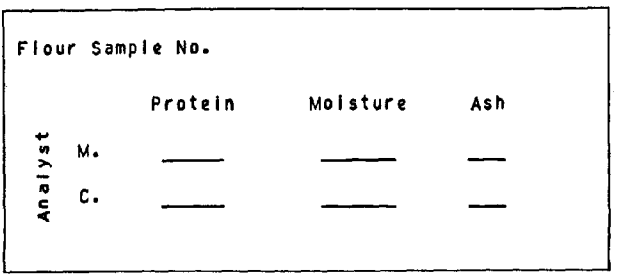
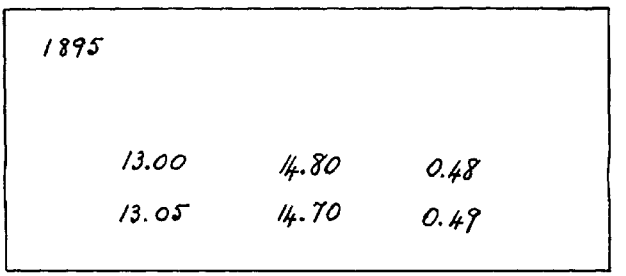


FIGURE 5

A Specimen Record Card From a Short Series Investigation (Data Identified by the "Key Card" of Figure 4)



punched with holes in positions specified to accord with a fixed plan of seriation for each variable. Sorting then proceeds by electrical contacts made through the holes as each card passes between wire brushes and a plate as electrodes, the position of the hole determining the destination of the card in the receiving mechanism. Interlined cards permitting of the inclusion of written records further enhance the scope of the system. The utility of the method in very large-scale investigations is well-nigh amazing, but of course every method may be expected to have its limita-

tions. The reader interested in the method must be asked to refer for further information to literature prepared by the manufacturers of the machinery, or to one of the special books prepared on the subject.

GRAPHICAL REPRESENTATION OF DATA

The orderly manner of change in the class frequencies as one progresses from the lowest to the highest class along the scale of measurement is reasonably well apparent from inspection in the first three of the foregoing illustrative tables. In Table 4, although N is very large in each case, definite obstruction to one's appreciation of the orderly nature of the change in frequency is introduced because of the disturbing effect of the irregular grouping. A much better appreciation of the nature of the orderliness may always be secured through graphical portrayal of the data. Before embarking on demonstration of this, it may be well to consider first some points with respect to graphs in general.

A statistical graph is a geometric drawing which aims to portray the nature of the relationship between two (or more) quantities which vary over a range of values. This portrayal of the relationship as a whole is effectively achieved through the *replacement of number by dimension.* A complete pattern of dimensions is readily assimilable by eye, whereas number comparisons are ordinarily appreciated only in pairs, and then as a result of processes in mental arithmetic.

The effectiveness of any graph in presenting a desired picture depends predominantly on the clarity of design underlying it. Simplicity of plan is a prime essential to that effectiveness. Boldness of execution is probably next in importance. Clearness of line, scaling, and lettering, together with adequacy of descriptive legends, impart a quality to graphs for which there is no substitute. Many a graph has been severely impaired in value through incorporating in it several ideas or so much detail that confusion results. Other graphs fail completely in giving the correct impression because of violation of the fundamental principle that the unit of dimension with respect to any axis must remain constant at all positions along that axis; otherwise dimension does not replace number in a simple manner.

As drawings on plane surfaces, graphs in their simpler forms are confined to picturing the relationship between two variables only, using for the scales of those variables the two directions at right angles in terms of which surfaces are ordinarily measured. These two directions are called axes. They are conventionally established in the so-called horizontal and vertical directions, or from left to right and from bottom to top of the page as it lies before one on a desk. The horizontal axis is known as the *axis of abscissas,* and the vertical the *axis of ordinates.* On

lines drawn in each of these two directions, scales may be measured off, and any point on the surface may then be referred to each of these scales, separately and jointly. Thus the association of two magnitudes may be expressed by a single point, and repeated associations of such paired magnitudes may be portrayed by a succession of points forming a line or by a dispersion of such points over the surface. The relationship which two quantities varying over their respective scales bear to one another may therefore be portrayed clearly in terms of a simple surface to give a unified picture of the whole.

All graphical dimensions are constructed from numbers. A graph cannot be any more accurate than the data from which it is constructed. It is only because a general view of problems involving many numbers may usually be secured much more readily from a suitable graph of the data that the graph is constructed. We may now look specifically into the problem of graphically portraying frequency distributions.

BAR DIAGRAMS AND HISTOGRAMS

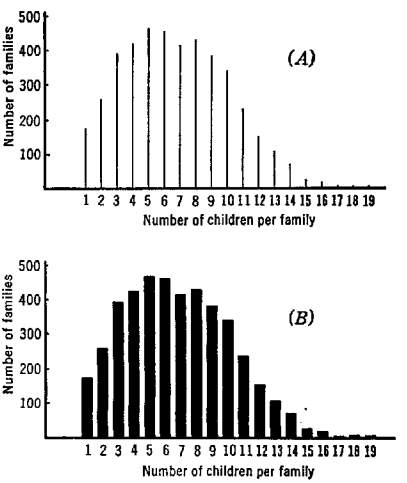
Graphical representation of distribution of frequency along the scale of a variable may readily be given, one axis (conventionally the abscissa) being used for the scale of measurement of the variable, frequency being portrayed in terms of the other axis. For any discrete variable which is not seriated into broad classes, frequency is assembled at successive integers on the scale of whole numbers. These integers may be ascribed as values to corresponding successive points on the abscissal scale. At these points ordinates may be erected, each of length proportional to the frequency at that point. In the upper panel of Fig. 6 a graph so drawn for the data of Table 1 is reproduced. Although technically correct, its deficiency in boldness because of the thin frequency lines is at once apparent. Even what one sees with these fine lines is really an exaggeration, for technically a line has no thickness. Technicalities may, however, be swept aside quite properly at times. Exaggeration of the frequency lines into prominent solid bars would obviously do much to add to the practical effectiveness of the picture, without which a graph has no value. This is done in the lower panel of Fig. 6. Graphs of this nature are commonly called *bar diagrams*.

In preparing to portray the frequency distribution of a continuous variable, one recognizes at once that a different situation exists. Frequency now is spread over a range; it is not to be represented by an ordinate at a point but by an ordinate moved over a range of the abscissal scale, that is, by a rectangular area. One may then ask: What does the scale of ordinates describe? The solution is really quite simple. If seriation be into classes of equal range, then all the rectangles designating

frequency will have equal bases and therefore the heights of the rectangles will be proportional to the frequencies *per class range of · · · units.* The ordinate scale is therefore one of frequency per · · · units on the scale of the variable. This seemingly technical refinement is most important to the correct graphical portrayal of a continuous frequency distribution. To ignore it will cause quite erroneous graphs to be drawn when the class ranges are irregular. Recognition here of a technicality

FIGURE 6

FREQUENCY DISTRIBUTION FOR NUMBER OF CHILDREN IN COMPLETED FAMILIES, NEW ZEALAND, 1916 *



is essential to the purpose of the graph itself. The *frequency* corresponding to each class range on the scale of measurement is portrayed by a rectangle erected vertically above that range as a base and of *area* proportional to the frequency.

The point giving the upper limit of the true range of one class in a continuous variable also specifies the lower limit of the next. Hence the graph will assume the form of a collateral array of rectangles on a common base line. The figure thus established is known among statisticians as a *histogram* (Greek: *histos*, a web or tissue; *gramma*, a thing written).

* For basic data, see Table 1, page 16.

In Fig. 2*a* the histogram is drawn for the distribution of length of left middle finger as classified on the right-hand side of Table 3. Taking any one class, for instance that of greatest frequency, the 691 individuals in it have values lying between 114.5 and 117.5 mm. This frequency pertains not to any one point, but to the range designated. Correct representation of the distribution must spread the frequency over the

TABLE 5a

BRAIN WEIGHTS AT AUTOPSY BY SEX

(Data of Retzius [3])

Class range in grams	Frequency	
	Males	Females
900- 949		1
950- 999		—
1,000–1,049		1
1,050–1,099		2
1,100–1,149	1	12
1,150–1,199	4	16
1,200–1,249	13	24
1,250–1,299	18	26
1,300–1,349	35	16
1,350–1,399	53	14
1,400–1,449	43	8
1,450–1,499	40	6
1,500–1,549	21	—
1,550–1,599	19	1
1,600–1,649	11	
1,650–1,699	3	
1,700–1,749	1	
Totals	262	127

range to which it belongs. The scale of ordinates in Fig. 2*a* therefore portrays the frequency per 3-mm. range of finger length. Since all classes in this histogram have the same range, the heights of the rectangles are proportional throughout to the class frequencies.

In a study of brain weight at autopsy for individuals of both sexes between 20 and 50 years of age, Retzius [3] has given the series of data reproduced as Table 5a herein. The reader may undertake to prepare

[3] A. Retzius. Ueber das Hingewicht der Schweden. Biol. Untersuchungen, N.F.Bd. 9, Cap. 4: 51-68. 1900.

histograms of the two distributions in an arrangement permitting of ready comparison. In this connection the following notes may be observed:

(1) Conversion of the frequencies within each sex to a percentage basis will eliminate the effect of unequal size of the two samples. The total actual frequency should be indicated against each histogram in all such transfers to a relative frequency scale.

(2) Confusion of the two histograms will result if one is superimposed on the other. This may be avoided suitably by using two panels, one directly below the other in relation to a common weight scale.

(3) Superimposed frequency diagrams of an approximate nature may be made by plotting each percentage frequency as a point above the appropriate class center value. When these points are joined in order an outline somewhat simulating a frequency curve is obtained. Such an outline is commonly known as a "frequency polygon." The two polygons may be distinguished easily by using full lines for one and broken lines for the other. The reader should determine for himself the reason why the frequency polygon is designated as an approximate representation of the frequency distribution.

TABLE 6

FREQUENCY DISTRIBUTION FOR AGE AT MARRIAGE OF BACHELORS AND SPINSTERS IN ENGLAND AND WALES FOR THE YEAR 1932

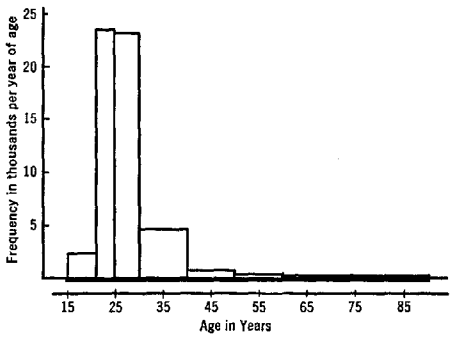
(1) Age group	(2) Frequency	(3) Class range (in years)	(4) Average frequency per year of age	(5) True age range
15–20	13,372	6	2,229	15–21
21–24	93,728	4	23,432	21–25
25–29	115,113	5	23,023	25–30
30–39	45,656	10	4,566	30–40
40–49	6,337	10	634	40–50
50–59	1,572	10	157	50–60
60–89	438	30	15	60–90

Sometimes it is inconvenient, or it is not possible with the data given, to preserve uniformity of class range in a histogram. Take, for instance, the data on number of marriages in England and Wales, 1932 (first marriages for both parties), given in the first two columns of Table 6.

Preparation of a histogram to portray the distribution of age at marriage as correctly as these limited data permit would necessitate first some such calculation as is given in columns (3) and (4) of the table. The histogram may then be drawn readily in terms of the figures in the last two columns of this table. It is presented as Fig. 7. The reader may test his grasp of this point by preparing histograms of the age distributions for two elements of the white population of Minnesota as recorded in Table 4. For this purpose the greatest age may be considered as 105, and the group of unknown age may be ignored.

FIGURE 7

FREQUENCY DISTRIBUTION FOR AGE AT MARRIAGE OF BACHELORS AND SPINSTERS IN ENGLAND AND WALES, 1932 *



THE FREQUENCY CURVE

The advantage of coarsening the grouping beyond the unit of measurement in order to portray graphically the form of a frequency distribution is quite pronounced whenever the total frequency is not many times larger than the number of increments of measurement between the smallest and greatest variate. Figure 8 illustrates the effect of coarsening the grouping progressively in smoothing out the histogram for a series of 402 new-born infant weights, measured originally to half ounces. This quite apparent smoothing as the class range is widened is due to the relatively smaller effect of errors of sampling on class frequencies as the latter are increased.

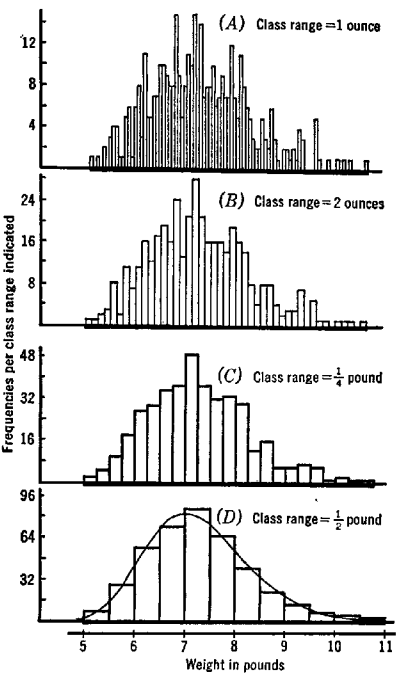
Let it now be assumed that it is possible to increase the number in the sample indefinitely, and that accordingly the unit of grouping of the material may be reduced to a much smaller range without loss of the

* For basic data, see Table 6, page 29.

general form of the distribution. Each rectangle in panel *D* of Fig. 8 would then be split into a succession of much finer rectangles, not all of the same height or frequency but changing in a systematic manner from one to the other as do the coarse groups in the sample of limited size. This is illustrated in Fig. 9, where development of the concept is given

FIGURE 8

FREQUENCY DISTRIBUTIONS FOR WEIGHT AT BIRTH OF 402 FEMALE INFANTS *



both logically and illogically in terms of the single broad class, 6.0 to 6.5 pounds. With an infinite number of such infant weights, and an increment of measurement of infinitesimal range, a frequency distribution bounded apically by a smooth curve must result. The type given by the flowing line in panel *D* of Fig. 8 would probably be obtained. This *frequency curve* has been fitted in accordance with critical statistical procedure, and does not represent free-hand graduation or reflect the

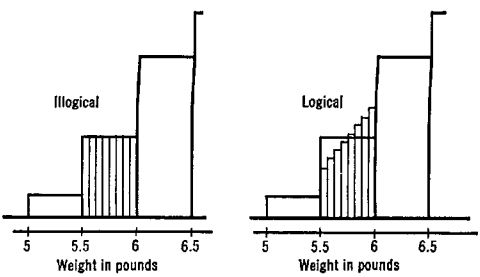
* For basic data, see Table 9, page 57.

influence of biological judgment arising from wider experience with this particular variable. As far as the given series of data alone may form a satisfactory basis for generalization, the curve depicts a statistical estimate of the form of distribution in the supply from which the sample has been drawn. It will be noted that this series of infant weights suggests a distinctly non-symmetrical distribution for this variable.

FIGURE 9

ILLOGICAL AND LOGICAL SUB-DIVISION OF FREQUENCY IN PASSING FROM COARSE TO FINER GROUPING IN A SYSTEM OF VERY LARGE FREQUENCY

(Illustrative Segments from Panel (*D*) of Figure 8)



SYMMETRY

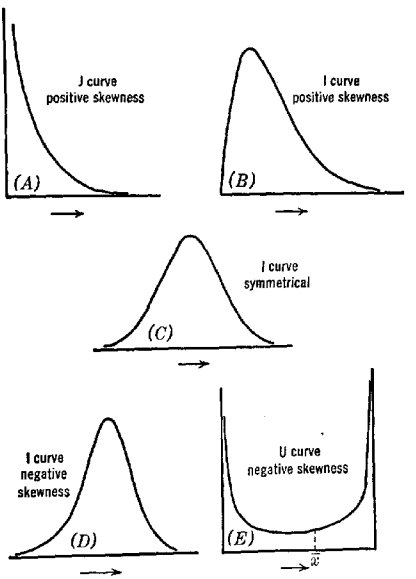
Frequency systems may primarily be divided into two main groups, the *symmetrical* and the *skew*. Figures 1 and 2 belong to the former group wherein positive and negative deviations from the central value have equal frequency. This balance of frequencies about the central value is not realized in the distribution curve of Fig. 8, which is designated as skew. One might say that in skew distributions one "tail" of the curve is more "drawn out" than the other. Skewness may occur in either direction; that is, it may be either positive or negative. These terms for direction of skewness are related to the direction in which the "long tail" extends. When that tail is toward larger values than the average it is called positive, and *vice versa*.

Skewness may be—although it rarely is in biology—of such magnitude as to change the form of the distribution from the **I**-shaped curves already presented, to forms designated for convenience as **J**- and skew **U**-shaped. The **I**-shaped curves are those in which the peak of frequency occurs between the limits of the distribution, whereas in the **J** forms a single peak occurs at one end of the distribution, and in the **U** forms a

peak occurs at each end of the distribution. There is, of course, a complete transition scheme for I-shaped curves from extreme skewness in one direction, through symmetry, to extreme skewness in the opposite direction. The J and skew U curves represent still more extreme departure from the central symmetrical I form. The sketches of Fig. 10 will serve to indicate the transition scheme.

FIGURE 10

ILLUSTRATING THE TRANSITION SCHEME IN SKEWNESS OF FREQUENCY DISTRIBUTIONS *



KURTOSIS

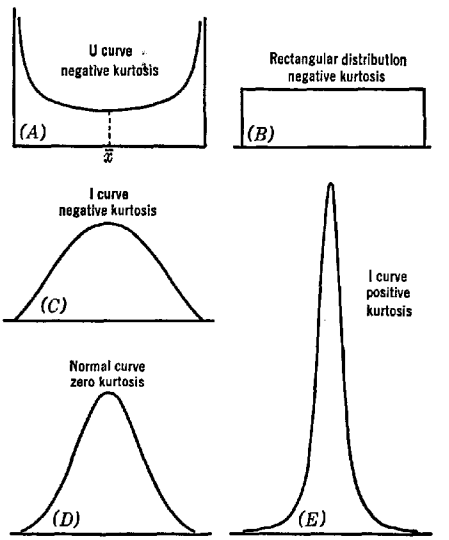
Symmetrical curves vary with respect to the relative concentration of the frequency at the center of the distribution. Small differences between curves in this regard are not so readily detectable as are those of skewness, although the characteristic is obvious when extreme forms are compared. This characteristic may be described as degree of peakedness. It is known technically as *kurtosis*, from the degree of curvature

* All curves drawn are from graduations of real variables.

at the peak. Positive (or lepto-) kurtosis, mesokurtosis (that of the *"law of error" or normal curve), and negative (or platy-) kurtosis* mean simply that the clustering at the center is respectively greater than, equal to, or less than that of the normal curve. The **U**-shaped symmetrical curve represents an extreme form of negative kurtosis, the rectangular distribution being somewhat less so, but this is too technical a subject for detailed consideration at this point. The sketches of Fig. 11 illustrate a wide range of kurtosis.

FIGURE 11

ILLUSTRATING THE TRANSITION SCHEME FOR KURTOSIS OF SYMMETRICAL FREQUENCY DISTRIBUTIONS



BIOLOGICAL VARIATION

Although variation may express itself in any one of a continuous array of forms departing from the central type in all degrees of skewness or kurtosis, biological variables for the most part do not differ very markedly from the "normal curve" or "law of error" form. While relatively small degrees of skewness and kurtosis are very common, large deviations from normality are the exception rather than the rule. This

has a most important bearing in simplifying statistical procedures for the biologist.

Quantitative description of the characteristic features of frequency distributions, which may be extended to any number of dimensions of space, comprises a basic task in statistical analysis. Readily comprehensible numerical quantities must be devised which will describe these features accurately, so that comparisons of them in different variables may be accomplished with ease and in an objective manner. Attention will now be given to the simpler of these problems.

CHAPTER 3

TYPICAL VALUES

Suppose a series of magnitudes, $x_1, x_2, x_3, \cdots, x_N$, all recognized as belonging to the same variable, to have been recorded in quantitative description. If all these individual values are identical, it is not necessary to determine mathematically an expression to represent them. The value required would be obvious by inspection. Any one value of x would be entirely representative of the whole series.

Complete concordance of all values of a variable is rarely, if ever, encountered in biological work. A situation of perfect concordance would (or should) immediately incite suspicion that the observations are biased. Indeed, in all fields of science, as the instruments for measurement become more refined it is increasingly apparent that things which may at first have seemed identical really differ among themselves. Under such conditions no single individual measurement can be accepted as being wholly representative of the entire sample. It becomes necessary to obtain a description of the series as a whole that will take into account all the measures available. A first step is to establish a *typical* value for the variate magnitudes—a single value which will be representative of the series of magnitudes in some specified and appropriate manner.

THE ARITHMETIC MEAN

The *average* or *mean* value is so familar as a typical magnitude that it seems superfluous to discuss its functions; and yet, like much that is commonplace, its philosophical implications rarely receive consideration. It provides a measure of type, the usefulness and appeal of which are attested by the extensive service which it has rendered since calculation first began. It is simple both to comprehend and to compute. It is perfectly impartial and unselective, for every observation enters into its magnitude without any trace of distortion. All who have employed it —and who, in fact, has not?—have in truth ascended to the first rung of the ladder of statistical analysis.

A very useful symbol for the mean value of a variable is provided by placing a bar above the letter designating the variable. If x is that variable, then the symbol for the mean, $\bar{x}$ (customarily verbalized as "x bar"), finds some preference over such alternatives as M_x (mean of x) or A_x (average of x) because of its avoidance of the subscript notation. In algebraic definition of $\bar{x}$, one may write,

$$\bar{x} = \frac{x_1 + x_2 + x_3 + \cdots + x_N}{N}$$
$$= \frac{\Sigma x}{N}. \qquad (1)$$

The Greek letter Σ (pronounced "sigma"), which is the equivalent of the English capital letter S, denotes summation for the entire series of values indicated by the symbols that immediately follow it. It is merely a mathematical shorthand symbol for continued addition of all values corresponding to the one given. When the equation is rearranged to the form

$$\bar{x} = \frac{x_1}{N} + \frac{x_2}{N} + \frac{x_3}{N} + \cdots + \frac{x_N}{N}$$
$$= \Sigma \frac{x}{N},$$

the mean appears as a sum compiled from one Nth part of every magnitude considered. Each individual in the series is therefore represented in the mean in proportion to its magnitude. This property defines the sense in which to the lay mind the mean logically typifies any series for which it is calculated.

Technically, equation (1) defines the *arithmetic mean* of a series. Two other types of average generated by the arithmetic mean on transformed scales and therefore appropriate to those special conditions justifying the transformation are the *geometric* and *harmonic means*, defined respectively by G and H in the equations:

$$\log G = \frac{\Sigma(\log x)}{N}, \quad \text{or} \quad G = \sqrt[N]{x_1 x_2 \cdots x_N};$$

$$H^{-1} = \frac{\Sigma(x^{-1})}{N}, \quad \text{or} \quad \frac{1}{H} = \frac{1}{N}\left[\frac{1}{x_1} + \frac{1}{x_2} + \cdots + \frac{1}{x_N}\right].$$

Since these statistics pertain to special types of analysis of relatively rare occurrence in biometric work, further discussion of them will not be given here. When the term *mean* is used without further designation, the arithmetic mean is always implied.

A fundamentally different approach to the problem of establishing a representative value for a series of variates may be made through selection of some particular variate or group of variates as typical of the whole. Two such selections came into rather common usage in the days when calculating machines were very much more of a luxury than they are now. These two descriptive values became more or less obvious as a result of organization of the data and do not depend primarily on calculation. They merit careful consideration at the present time, not as short-cut approximations to the mean for which purpose they appear to have been widely used, but as quantities descriptive of type in different senses from the mean. One may perhaps more readily grasp the descriptive value of these quantities through derivation of them directly from actual data.

THE MEDIAN

In a study of the erythrocyte count (red blood cell count) for normal men, Haden [1] gives findings for 40 individuals. Arranged in order of magnitude, these counts were as follows:

4.27, 4.32, 4.40, 4.52, 4.56, 4.58, 4.64, 4.70, 4.72, 4.73,
4.80, 4.80, 4.80, 4.80, 4.84, 4.87, 4.89, 4.93, 4.97, 4.98,
4.99, 5.00, 5.02, 5.05, 5.09, 5.09, 5.10, 5.15, 5.16, 5.20,
5.20, 5.20, 5.26, 5.28, 5.36, 5.46, 5.49, 5.50, 5.57, 5.62,
millions per cubic millimeter.[2]

This arrangement of a series of values in order of magnitude is known as *ranking*. It enables one to determine quickly how many individuals fall below and above any given magnitude. The problem is to select a magnitude which is representative of the set. One may endeavor to do this with the idea of a central value in mind. Since the above series has an even number of variates, there is no "middlemost" individual. However, 20 individuals have a count not exceeding 4.98, and 20 have a count not less than 4.99. The mid-point between these two magnitudes may logically be selected as the middlemost magnitude. It is known technically as the *median* value. Had N been odd, then there would have been a middlemost individual, giving a unique solution to the problem of ascertaining the value above and below which an equal number of observations occur. Of course the situation would become complicated

[1] R. L. Haden. Accurate criteria for differentiating anemias. Arch. Int. Med., 31: 766. 1923.

[2] Note that the count is made under a microscope in a volume equivalent usually to only one ten-thousandth of a cubic millimeter of actual blood. The result is then multiplied by 10,000 and the count reported as per cubic millimeter, hence the jumps by a smallest increment of 10,000 in the figures recorded above.

again if several variates had that same magnitude, but we need not pause to argue this point. The median may be accepted as being representative in the sense that it divides the total frequency into halves; it is central with respect to the distribution of the total frequency along the scale.

As the central value with respect to frequency distribution, the median has the peculiar property of being unchanged by alteration of any or all of the other variate magnitudes in the series, provided that the *sign* of deviation from the median remains unchanged in each case. Stated inversely, the numerical magnitudes of the individual deviations from the median in a sample do not influence the median for that sample. Thus the median is often quite useful as a measure of "central tendency" in a sample where it is desired to suppress the influence of the extreme or unusual variate magnitudes. It is a magnitude representative of mediocrity, serving to eliminate the influence of the geniuses and the morons for the student of educational statistics, or the influence of "epidemics" and periods of unusual good health for the vital statistician, and so forth.

THE MODAL VALUE

The organization of a series of data by ranking the variates places identical magnitudes together and automatically draws attention to the relative frequencies of occurrence of the observed values. It is a logical stepping stone to the form of organization which we have already considered under the title of seriation. Just as ranking naturally centers attention on the number of observations above and below any specified point, so seriation emphasizes the importance of successive values with respect to frequency of occurrence. The class of greatest frequency naturally draws attention on this account, *if* the class ranges are uniform. It is called the *modal* or "most fashionable" class, the class center being commonly designated as the *modal value*. This provides immediately a value which is typical of the observations in a new and quite appropriate sense.

It is well to recognize at once that establishment of the modal value is fundamentally associated with a reasonably satisfactory definition of this point in the appropriate law of frequency distribution for the variable considered. If any seriation does not lead to a reasonably smooth frequency distribution, the question arises at once of the acceptability of the indicated modal value. Now "smoothness" of a histogram is a function of size of sample and the unit of classification. This unit is chosen quite arbitrarily, and the indicated modal value may be shifted substantially according to this selection, even when N is quite large.

The difficulties of selecting the appropriate modal value by seriation of a small sample are often overwhelming. Let us look for a moment at the series of 40 erythrocyte counts above. The count of 4.80 occurs four times and is the most frequent individual value as the data are given. But what would one do about selecting the modal value if a recheck of the counts showed an error, one of the 4.80 values being changed, say, to 4.81? In that case both of the counts 4.80 and 5.20 would occur with a frequency of 3, giving two modal values rather widely separated. Coarser grouping may be resorted to, but the indicated modal value will be found to vary around 5.00 according to the seriation scheme adopted. Study of the histograms for the same series of observations in the successive panels of Fig. 8[3] will quickly substantiate the general principle. As applied to a small sample, the modal value has rather poor descriptive capacity. From what has been said in the preceding chapter it is easy to see that this deficiency will diminish as the size of the sample is increased and its frequency distribution accordingly reflects more clearly the supply form.

DESIRABLE QUALITIES OF A DESCRIPTIVE STATISTIC

These preliminary definitions of the mean, median, and modal values as independent and differing descriptions of typical value raise the question of the relative quality of alternative descriptions. It has been stated that quantitative description of the characteristic features of sets of data comprises a basic task in statistical analysis. Numerical quantities must be devised which will describe these features accurately, so that comparisons of them in different variables may be readily accomplished in an objective manner. The term *statistic* is now commonly being applied to any summarizing quantity such as a mean, median, *et cetera*, derived from and descriptive of a set of variates. Using the term in this sense we may now well inquire as to the desirable qualities of good descriptive statistics in general. The following list is suggested:

(1) They should be based upon every variate in the sample. Assuming the basic data to be accurate, surely each measurement is just as important as any other measurement in defining the variable.

(2) They should be capable of algebraic definition. Algebra may be considered as providing the vehicle of symbolic definition for the development of logical thought sequences applicable to magnitudes differing by finite amounts. The utility of any statistic as a factor in such logical development of thought will depend on the generality of its algebraic definition.

[3] *Vide* page 31.

(3) They should be readily comprehensible with respect to each descriptive function.

(4) They should be susceptible as far as possible of rapid determination.

(5) They should yield highly consistent values when applied to samples randomly drawn from the same supply.

While the mean fulfills every one of these conditions remarkably well, the same cannot be said of the modal and median values with equal validity. The latter are not capable of algebraic definition in general terms, nor do they depend directly for their values on the magnitudes of every variate in the series. Also, the consistency of means in random samples from the same supply is found to be considerably greater than that of median or modal values. If these three statistics are to be considered as alternatives in description, there can be no question as to the superiority of the mean. It should not be overlooked, however, that each has arisen as the result of a different point of view in approaching the problem of establishing a representative value. The problems in which they assume quite different magnitudes must be regarded as problems in which they are not alternative, each having its own descriptive function.

"CENTERING VALUES" OF FREQUENCY CURVES

The preceding syntheses of the mean, median, and modal values as representations of typical magnitude referred to the law of frequency distribution only in the case of the modal value. The descriptive quality of the modal value was found to be very poor with small samples, but increased as refinement of seriation could proceed without loss of the form of the frequency distribution, that is, as the sample increased in size. It is only in the limit of a frequency curve being reached that the point of greatest frequency becomes independent of the various seriations to which the data may be subjected. It was urged by Karl Pearson that the term *mode* be reserved for the scale value at which the greatest ordinate occurs for the fitted frequency curve. Our previous use of the variant term *modal class* for the range within which maximum frequency occurs in a seriation was made with this distinction in mind.

When a histogram is graduated by a frequency curve, no attempt is made to force the mode of the curve to agree with the modal class of the histogram. The constants of the equation used fix the point at which the maximum ordinate will occur. These constants are determined from general properties of the observed data and are essentially independent

of any seriation made. The curve given in panel (*D*) of Fig. 8[4] must logically graduate each of the seriations given equally well. The various modal values of the several seriations contrast sharply with the fixed mode of the curve at 7.00 pounds.

The mean and median of a frequency curve are established precisely in accordance with the definitions already drawn for the variates. The median becomes the point on the scale of measurement at which the ordinate cuts the frequency curve into two segments of equal area, the latter—as we have sought to make clear—defining the frequency. It will be shown in Chapter 5 that the algebraic definition of the mean (equation 1) fixes it as the "point of balance" of a frequency system. That is, if a frequency distribution bounded apically by a curve is cut out of cardboard of uniform thickness and is balanced horizontally about an ordinate, then that ordinate will emanate from the base line at the mean value. It follows directly then that positive and negative deviations of the variates about any point in the scale of measurement of a variable balance one another *in toto* only when that point is taken at the mean of the distribution. This defines another attractive quality of the mean as a representative value.

For symmetrical frequency curves, the median, the mode, and the mean all coincide, as must surely be obvious. In a skew curve this is not so. The point of balance (the mean) must necessarily be drawn increasingly away from the point of greatest frequency (the mode) toward the long "tail" of the distribution as skewness intensifies. The amount of separation between them accordingly forms an excellent basis for quantitative description of the degree of skewness present in a frequency distribution.

The median of non-symmetrical J-shaped curves falls between the mode and the mean, somewhat closer to the latter than the former. For small degrees of skewness it holds very closely to the approximating equation

$$(\text{mode} - \text{mean}) (=) \; 3\,(\text{median} - \text{mean}).$$

Solving this equation to provide quickly an estimate of the mode after mean and median have been determined from a set of observations is perhaps acceptable as a rough approximation only if the sample is a large one and skewness is not very marked. Rather extensive calculation is involved in accurate determination of the mode, for the equation to the graduating curve must be fully determined first. Because of this there is temptation to use the above approximation without proper regard to its limitations, a temptation to be avoided, of course.

[4] *Vide* page 31.

In that which follows in this book we shall assume, as indeed we must to progress satisfactorily, that simple computations present no obstacle. It will be shown in the following section that a powerful device for the simplification of calculations may be mastered easily, placing the bulk of statistical work within the scope of elementary mental arithmetic. In view of this and the reasonable approach to symmetry commonly found in biological variables, the natural advantages of the mean as a typical value and reference point for the measurement of variation will cause its adoption in preference to the median for practically all work.

CODING AS AN AID IN COMPUTATION

The mean must be determined by calculation. Algebraically, this seems a most simple procedure, and indeed it is, provided

(1) that the number of variates to be summed is not so large as to cause fatigue when handled individually, and

(2) that the variate magnitudes are numbers of very few digits.

If either of these difficulties is present, they may readily be circumvented. The first may be avoided by seriation to form a frequency table. The second is very simply overcome by *coding*, which consists of replacing the true variate magnitudes by a simpler set of numbers according to a systematic scheme. The original scale of measurement may be replaced by an arbitrarily chosen scale of the simplest possible form in accordance with a well-known mathematical principle. All computations may be made in terms of the arbitrary scale and the results quickly transferred to their corresponding original scale values.

The objectives of seriation have already received attention in these pages. It may not be amiss at this point, however, to consider certain procedural details as they relate to the establishment of the computational table. Solving the following practical problem will serve as a medium for development of these points, as well as others incidental to the subject of coding.

In routine analysis of the protein content of grain arriving at a port, 1,937 determinations are made available. The protein content of each sample is expressed on a percentage basis to the nearest 0.05 per cent. The lowest value is 10.30 per cent, and the largest is 17.15 per cent, practically all intermediate values by increments of 0.05 occurring. How may the computation of the mean for such a series of values be most advantageously simplified?

The very large number of variates to be considered makes seriation a practical necessity. Since the increment of measurement is 0.05, there will be 138 possible values between the extreme limits of 10.30 and 17.15. Grouping into classes of sufficiently wide but uniform range to reduce

the number of classes to about 12–16 would be desirable. A class range of 0.5 per cent protein would give approximately 14 classes. That range is so convenient for seriation purposes that it may well be chosen.

"At what value should the beginning class boundary be set?" The reply to this question must be that, like selection of the class range, this matter remains entirely at the discretion of the individual. Since the class centers become the reference values after the seriation is made, some workers would aim to choose the boundaries so that the centers become simple numbers, like the even and half percentages. This selection would place the class boundaries at values ending in .25 and .75, and the question would arise as to what to do with the large numbers of variates which fall precisely on these boundaries. Having the true class range an *odd* multiple of the increment of measurement avoids this difficulty, but, having due regard to ease of seriation, that would be inconvenient here, so we shall proceed with the suggested class range of 10 increments of measurement.

Having as simple a pattern of numbers as possible for the apparent class boundaries greatly facilitates seriation. Thus apparent class ranges of 10.0 – 10.45, 10.50 – 10.95, 11.0 – 11.45, *et cetera*, will lead to quicker seriation than possibly any other scheme with a true class range of 0.5 unit. We shall see later that it is quite immaterial to the calculations what the complexity of the class centers may happen to be, so we shall seriate according to this plan. Table 7 presents in its first two columns the frequency distribution found for the 1,937 variates in the series.

Since the measurements were recorded to the "nearest" scale graduation, the mid-point of the apparent range for each class is identical with the center of the true class range. These centers are entered as column (3) of the table. For computational purposes they will replace the ranges, the effect of the seriation being to represent the data as if they had been recorded by increments of 0.5 per cent instead of 0.05.

Proceeding now to compute the mean value for this distribution, one must first multiply each class center, x, by the number of individuals in the class, f, to secure the sum of all variate magnitudes within the class. Column (4) contains these products, the grand total of the column being Σx for the entire distribution. By dividing this total by 1,937, the mean protein content is found to be 13.233 per cent.

The reader who has taken the trouble to verify the calculation of the mean in Table 7, if he did not use a machine for multiplications, will have been impressed by the onerous mental arithmetic imposed by the five-digit numbers for the class centers. It is the particular function of coding to reduce this labor to a minimum through replacement of the

existing set of class center or x values by a simpler set. In practice, coding is a most direct and elementary procedure, used in one form or another on occasion by all. It would seem advantageous first of all to trace the underlying steps with respect to a very simple illustration and then give full generality to the principle in algebraic form.

TABLE 7

A FREQUENCY DISTRIBUTION OF PROTEIN ANALYSES, WITH CALCULATIONS FOR THE MEAN VALUE

(Original determinations to the nearest 0.05 per cent protein)

(1) Apparent class range	(2) Frequency f	(3) Class center x	(4) Class total fx
10.0–10.45	4	10.225	40.900
10.5–10.95	18	10.725	193.050
11.0–11.45	65	11.225	729.625
11.5–11.95	157	11.725	1,840.825
12.0–12.45	270	12.225	3,300.750
12.5–12.95	327	12.725	4,161.075
13.0–13.45	334	13.225	4,417.150
13.5–13.95	276	13.725	3,788.100
14.0–14.45	212	14.225	3,015.700
14.5–14.95	141	14.725	2,076.225
15.0–15.45	79	15.225	1,202.775
15.5–15.95	31	15.725	487.475
16.0–16.45	14	16.225	227.150
16.5–16.95	7	16.725	117.075
17.0–17.45	2	17.225	34.450
	1937		25,632.325

$\Sigma x = 25{,}632.325.$ $N = 1{,}937.$

$\bar{x} = 13.233.$

Probably any science student asked to give the average of the three numbers 19.01, 19.06, and 19.08 would immediately recognize that, since 19.0 is common to all, the average must be 19.0 followed by the average of 1, 6, and 8. He would be coding the data in so doing, finding the average on a new scale, then decoding his answer to get 19.05. Let us follow the steps in detail.

$$\frac{19.01 + 19.06 + 19.08}{3} = \frac{19 + 0.01 + 19 + 0.06 + 19 + 0.08}{3}$$

$$= \frac{3(19)}{3} + \frac{0.01 + 0.06 + 0.08}{3}$$

$$= 19 + \frac{0.01(1 + 6 + 8)}{3}$$

$$= 19 + 0.01\left(\frac{15}{3}\right)$$

$$= 19 + 0.01(5)$$

$$= 19.05.$$

"But," he might protest, "I didn't go through all that manipulation!" Perhaps what he should have said was that he did not realize that his short cuts were based on just that type of manipulation.

Let us now generalize these steps. Let x designate the original scale values, a be the constant 19, and b the constant 0.01. The numbers actually averaged (1, 6, and 8) were derived from the x values by subtracting a from each, then dividing the remainders by b. Designating the resulting code values by x, we have

$$\mathrm{x} = \frac{x - a}{b},$$

or

$$x = a + b\mathrm{x}.$$

Now

$$\bar{x} = \frac{\Sigma x}{N}$$

$$= \frac{\Sigma(a + b\mathrm{x})}{N}$$

$$= \frac{\Sigma a}{N} + \frac{\Sigma b\mathrm{x}}{N}$$

$$= a + b\frac{\Sigma \mathrm{x}}{N}$$

$$= a + b\bar{\mathrm{x}}. \tag{3}$$

Therefore the mean value on the original scale of measurement may be found by securing the mean value in terms of a code scale and then decoding.

The reader who does not recall the algebraic rules of rearrangement about the summation sign, Σ, may easily verify each step of this deriva-

tion by replacing Σx in the first line by its equivalent full form, $x_1 + x_2 + \cdots + x_N$. He will in this manner establish the following rules for himself:

(1) When summation of an expression involving terms joined by plus or minus signs is called for, the summation may be made by individual terms.

$$\Sigma(a + bx - y) = \Sigma a + \Sigma bx - \Sigma y.$$

(2) Summation of a constant is equal to N times that constant, if the summation is over N elements.

$$\Sigma a = Na.$$

(3) Summation of the product of a constant and a variable is equal to the product of the constant and the summation of the variable.

$$\Sigma bx = b\Sigma x.$$

Note also:

(4) Summation of the product of two variable magnitudes cannot be rearranged.

We may now return to the frequency distribution of the protein determinations, reproduced anew in the first two columns of Table 8. Can we reach a simple scale to replace the awkward x values by coding according to the foregoing scheme? Very simply! First take one of the values of x and use it as the constant a. Two such selections are used in the table: (1) a_1 is taken as the lowest value of x; (2) a_2 is taken as a value near the center of the series, usually the modal class value. The respective series of residuals, $x - a$, are given as columns (3) and (5). Now take as the constant b the interval on the residual scale, which is of course the true class range on the original scale. Dividing the residuals by b in each series gives the code scales x_1 and x_2 of columns (4) and (6). The contrast of the simplicity of these code scales with the original class centers surely needs no emphasis.

Statistical organization of the data has made possible the expression of magnitude for the variable in the simplest possible form for computational purposes. The new scales, although arbitrarily determined, do not lack descriptive value in the least. For instance, 8 on scale x_1 means 8 steps of 0.5 above the value 10.225, and -3 on scale x_2 means 3 steps of 0.5 below the value 13.225.

The mean value of the series on each new scale may now be secured with easy mental arithmetic. Proceeding as on Table 7, the sums of variates within classes, $f\mathrm{x}_1$ and $f\mathrm{x}_2$, are given in columns (7) and (8) of

Table 8. Dividing the grand totals in each case by the total number of cases (1,937) gives $\bar{x}_1 = 6.016$, and $\bar{x}_2 = 0.016$.

TABLE 8

ILLUSTRATION OF THE PRINCIPLE OF CODING BY LINEAR TRANSFORMATION OF SCALE FOR A FREQUENCY DISTRIBUTION OF UNIFORM CLASS RANGE

(Data of Table 7)

Class centers	Frequency	Coding schemes (1)		Coding schemes (2)		Class totals	
		$a_1 = 10.225$	$b = 0.5$	$a_2 = 13.225$	$b = 0.5$		
x (1)	f (2)	$x - a_1$ (3)	x_1 (4)	$x - a_2$ (5)	x_2 (6)	fx_1 (7)	fx_2 (8)
10.225	4	0.0	0	−3.0	−6	0	− 24
10.725	18	0.5	1	−2.5	−5	18	− 90
11.225	65	1.0	2	−2.0	−4	130	−260
11.725	157	1.5	3	−1.5	−3	471	−471
12.225	270	2.0	4	−1.0	−2	1,080	−540
12.725	327	2.5	5	−0.5	−1	1,635	−327 −1712
13.225	334	3.0	6	0.0	0	2,004	0
13.725	276	3.5	7	0.5	1	1,932	276
14.225	212	4.0	8	1.0	2	1,696	424
14.725	141	4.5	9	1.5	3	1,269	423
15.225	79	5.0	10	2.0	4	790	316
15.725	*31*	*5.5*	*11*	*2.5*	*5*	*341*	*155*
16.225	*14*	*6.0*	*12*	*3.0*	*6*	168	*84*
16.725	7	6.5	13	3.5	7	91	49
17.225	2	7.0	14	4.0	8	28	16
							+1743
Totals	1,937					11,653	+31
Means						6.016	0.016

According to these values the mean is 6.016 steps of 0.5 above 10.225, or 0.016 step of 0.5 above 13.225. In algebraic form,

$$\bar{x} = a + b\bar{x}.$$

That is,

$$\bar{x} = 10.225 + 0.5(6.016)$$
$$= 13.233;$$

or again,

$$\bar{x} = 13.225 + 0.5(0.016)$$
$$= 13.233.$$

As an exercise the reader may calculate $\bar{x}$ for the finger-length series given in Table 3,[5] first using the true scale without coding, then using one or both of the coding schemes just given. If Table 2 is now referred to,[6] it will be observed that subtraction of a constant a is all that is necessary to yield a desired code scale in that case. This amounts to the constant b being taken as unity, so that $\bar{x}$ is equal to a plus $\bar{\mathrm{x}}$. In Table 1[7] the original scale is itself of the simple type; any coding, if undertaken, would be confined to taking a equal to 7, say, so that each multiplication would involve a one-digit figure.

The reader will quickly perceive that the columns (3) and (5) of Table 8 are quite unnecessary in routine coding of the x scale in a frequency distribution table, provided that the seriation is uniform. Whichever type of code scale is desired may be written in directly and the values then recorded for the constants a and b involved. Thus, a is the value on the original scale corresponding to zero on the code scale, and b is the true class range (or interval between class centers) on the original scale.

The advantage of taking a near the middle of the series of class centers, giving the smallest values possible on the code scale, is offset to some workers by the resulting introduction of minus signs into the computations. When the number of classes is much in excess of 12, however, a code scale with zero toward the center may facilitate mental arithmetic quite considerably. Some workers using code scales with negative values choose to make a guess at the mean on the original scale, then take a as the x value immediately below this. If the guess has been good, $\bar{\mathrm{x}}$ will be a very small but positive value. Should $\bar{\mathrm{x}}$ prove to be negative in any case, however, it must be remembered that its sign is essential to it in any further calculation. Thus,

$$a + b(-\bar{\mathrm{x}}) = a - b\bar{\mathrm{x}}.$$

A seriation involving irregular grouping often makes full coding according to the above scheme of very little if any value. However, the first step alone of subtracting a suitable value a may help computations considerably. This abbreviated coding is often used when seriation of data is not worth while and a constant may be subtracted readily by mental arithmetic as each variate magnitude is handled. In such cases, a is best taken as a round number somewhat less than the lowest variate so that mental subtractions will give immediately obvious residuals.

[5] *Vide* page 19. [6] *Vide* page 17. [7] *Vide* page 16.

CHAPTER 4

THE MEASUREMENT OF VARIATION

Francis Galton wrote, in his classical book "Natural Inheritance"

> It is difficult to understand why statisticians commonly limit their inquiries to Averages, and do not revel in more comprehensive views. Their souls seem as dull to the charm of variety as that of the native of one of our flat English counties, whose retrospect of Switzerland was that, if its mountains could be thrown into its lakes, two nuisances would be got rid of at once. An Average is but a solitary fact, whereas if a single other fact be added to it, an entire Normal Scheme, which nearly corresponds to the observed one, starts potentially into existence.

One would have little ground for castigating statisticians today in the words used by Galton nearly 50 years ago, and yet his eloquent portrayal of the loss sustained by failure to describe variation has imparted lasting quality to that once-merited criticism. To attempt to describe a series of magnitudes solely in terms of some typical value and thus sacrifice all knowledge of their variability is tantamount to obliteration of a most crucial feature of the character being measured. As a basis of generalization in science, each sample statistic such as a mean has its degree of trustworthiness which depends most fundamentally on the amount of variation present in the recorded values from which it is computed. How enriched an average becomes when a measure of its sampling stability is added to it! The biologist should certainly not fail to measure the variation he encounters in his quantitative observations. Thus may he learn to defend himself against both the embarrassments of overhasty inference and the losses of overwrought caution.

RANGE

Reference to the amount of variation shown by the graph of a frequency distribution would incite contemplation of the "amount of spread" of the distribution as it may be read in terms of the base scale. The *range of variation*, or difference between the minimum and maximum values in any seriation, might well be offered as a first measure of variation. In terms of a histogram this is a finite measure, but if attention is given to frequency curves the appeal of the range fades away like the

frequencies at each end of the distribution. In the supply portrayed by the curve, just where does frequency stop? In most cases there is nothing harder to determine with real satisfaction than the answer to this question.

The range of any finite sample is of course quite definite. But what is one to do about the outlying case which appears so commonly to plague the investigator with conjecture as to its "really belonging" with the rest of the individuals in the sample? Exclusion of that one measurement may change the range markedly. The importance which the extreme individuals assume in this portrayal of variation is overwhelming. Is it of no consequence at all where the intermediate variates fall? One faces many such disturbing problems in careful analysis of the descriptive quality of range as a measure of variation. The need for a fundamentally sounder approach to the general problem soon becomes manifest.

INDIVIDUAL DIFFERENCES

Variations are observed through recognition of differences between individuals. Description of variation may on occasion be undertaken in that way. Variation in the characteristics of the members of a family may well be, indeed generally is, measured by direct comparison of each individual with every other. The justification of this method is due chiefly to two factors: first, a personal interest existing in each member as an entity; and second, most families are of distinctly limited number and the comparisons do not become onerous. The existence of such factors as these represents a very special situation when one turns to study of variation systems in general.

Algebraic analysis readily shows that among N individuals the number of different comparisons possible is equal to $\frac{N}{2}(N-1)$. When there are but 20 variates, 190 individual differences appear, and 100 variates would yield 4,950 comparisons. The difference between the minimum and maximum variates—the range—is of course just one of the set in each case. To attempt summary of individual differences by considering all possible pairs is obviously out of the question.

It may well have occurred already to the reader that the set of $N-1$ deviations of one selected variate from all others in any series must give all the information incorporated in the much larger set of $\frac{N}{2}(N-1)$ individual differences. The scientist is not interested in the latter values as such, but in the amount of variation characterizing the N variates. But if the deviations from some one variate are taken,

which one would be most appropriately chosen as this reference standard? This is a troublesome question. No one variate is clearly the best. Indeed, are not all equal in importance?

After all, the value to be chosen for this attack on measuring variation is to be a standard of reference. Need it be one of the N individuals of the series? What could provide a better standard than a value representative of all; for instance, one of the typical values discussed in the preceding chapter. What would be more logical than that, having established a typical magnitude, one should turn attention to variation in terms of deviations of the individuals from that magnitude? Consideration of the suitability of the three typical values for the purpose quickly eliminates the modal value, leaving the issue of choosing between the median and the mean. Of these, only the mean is fully representative of all variate magnitudes, for the median is just central with respect to total frequency. One may thus be guided to the mean as the standard of reference.

DEVIATION FROM TYPE

Just as the variates, $x_1, x_2, x_3, \cdots, x_N$, formed the basis for the measurement of type, so the quantities

$$(x_1 - \bar{x}),\ (x_2 - \bar{x}),\ (x_3 - \bar{x}), \cdots, (x_N - \bar{x})$$

provide a starting point for the measurement of variation or deviation from type. For brevity, one may designate the deviation of each variate from the *mean* of its series as a *deviate*. The task is to summarize these deviates into a single value that has, as far as possible, the qualities stipulated in the previous chapter for a good descriptive statistic. If these conditions are adequately fulfilled it may serve as our description of the amount of variation in any series.

There is no simpler form of summary than that provided by determining the average value. Would the average of the deviates conform with requirements for the summary of variation? Unfortunately the answer must be in the negative. Since the mean is the "balancing point" of the variates, the sum of the positive deviations from the mean must equal that of the negative deviates, regardless of the amount of variation present in the series. Algebraically,

$$\frac{\Sigma(x - \bar{x})}{N} = \frac{\Sigma x}{N} - \frac{\Sigma \bar{x}}{N} = \bar{x} - \bar{x} = 0.$$

This difficulty of positive and negative deviations which just balance one another in the sum may be removed in either of two simple ways.

First, the deviations may be considered without regard to their signs; all deviates may arbitrarily be considered as positive. Second, the deviations may be squared, or indeed raised to any even power, to give quantities of positive sign throughout.

The average of the deviates taken without regard to their signs has been used somewhat as a measure of variation. It is often called the *average deviation*. This quantity is very easy of comprehension and would appear to summarize the deviations satisfactorily. However, it is not widely useful as a measure of variation in more advanced statistical work. This is due to the great difficulty in solving algebraic equations involving quantities whose signs are to be ignored. Knowledge concerning such procedures is comparatively scarce, and the statistical worker would seriously limit his mathematical field, cutting himself adrift at once from a vast amount of that very logic which it is his desire to turn to his advantage, if he arbitrarily ruled all deviates as positive quantities. Consideration of the positive and negative deviations in two separated groups would also be unsatisfactory, since such a procedure would entail the use of two magnitudes to express the one idea of amount of variation, and would involve any further reasoning based on that idea with considerable complexity.

It is a derived property of numbers that the product of two negative quantities must itself be a positive number. All have learned the elementary rule of algebra flowing from this, that the *square* of any number is always positive, regardless of the sign of that number. The expedient of ridding the problem of the measurement of variation from a vexatious difficulty imposed by the signs of the deviates, through using the squared values rather than the absolute deviations, is very appealing in its mathematical simplicity. Very early in the quantitative measurement of variation this expedient was resorted to, and the results have proved so useful that the general statistical practice today is to measure variation as a mathematical function of the squared deviates, $(x - \bar{x})^2$. Summing these quantities and dividing by N, one will secure an average value typifying the squared deviations. Symbolizing this average by m_2, for it is commonly known as the *second-moment coefficient*, the equation for the mean squared deviate becomes

$$m_2 = \frac{\Sigma(x - \bar{x})^2}{N}. \tag{1}$$

In the complete absence of variation, m_2 must obviously be zero. When only a little variation is present, the values of $(x - \bar{x})^2$ are individually small and therefore their average is small. As the values of x

differ increasingly among themselves, m_2 must rise rapidly in magnitude. Also, the mean squared deviate is fully representative of all members of the series in that it takes into account every variate in the same manner.

As a *numerical* quantity this mean squared deviate seems to fulfill those conditions which are logically to be imposed for the measurement of deviation from type. However, variation is in our everyday thinking a matter of simple differences, not of squared differences. Squaring was resorted to for the purely mathematical reason of eliminating signs. The immediate objective being accomplished, one may now reverse the move by extracting the square root of m_2 so as to return to the original scale of measurement. These adjustments of squaring the deviates, then taking the square root of their mean, comprise transformations to and from a scale of squares made for convenience in hurdling an obstacle. The resultant quantity is a magnitude on the original scale of measurement and one directly proportional to the amount of variation present. It was called the *standard deviation* by Karl Pearson in 1894, and indeed it has since then become in fact the standard measure of variation.

A statistic of tremendous utility, the standard deviation lends itself admirably to use in advanced statistical work. Every effort given immediately to comprehend its nature and meaning as a measure of variation will be well repaid. Symbolizing it by s, the fundamental definitive equation may be written in the form

$$s_x = \sqrt{\frac{\Sigma(x - \bar{x})^2}{N}}. \qquad (2)$$

CALCULATION PROCEDURE FOR THE STANDARD DEVIATION

The Transfer from Deviates to Variates

The basic definition of the standard deviation is shown at a glance by equation (2) above. This formula is ill suited for computational purposes, however. The deviations from the mean when numerically expressed are generally cumbersome, and squaring them is an onerous procedure providing considerable opportunity for mistakes. Harris pointed out in 1910 that a simple transfer from deviates to variates very greatly reduces the work involved. The rearrangement may be shown algebraically as follows:

$$s_x^2 = \frac{\Sigma(x - \bar{x})^2}{N}$$

$$= \frac{\Sigma(x^2 - 2x\bar{x} + \bar{x}^2)}{N}$$

$$= \frac{\Sigma x^2}{N} - \frac{\Sigma 2x\bar{x}}{N} + \frac{\Sigma \bar{x}^2}{N}$$

$$= \frac{\Sigma x^2}{N} - 2\bar{x}\frac{\Sigma x}{N} + \bar{x}^2.$$

But $$\frac{\Sigma x}{N} = \bar{x}.$$

Therefore $$s_x^2 = \frac{\Sigma x^2}{N} - \bar{x}^2, \tag{3}$$

and $$s_x = \sqrt{\frac{\Sigma x^2}{N} - \bar{x}^2}. \tag{4}$$

Thus the standard deviation of a series of magnitudes may be derived expeditiously from the means of the first and second powers of the variates; the individual deviates need not be computed.

When a calculating machine is available and N is not too large (not in excess of 100, say), Σx and Σx^2 may be accumulated directly on the machine without previous seriation of the data. The mean and mean square may then be determined, the square root of the difference between them giving the standard deviation. Applying this procedure to the erythrocyte counts given by Haden for 40 normal men,[1] the following record sheet of computations was secured.

Series: Haden (1)

x = red blood cell count

$N = 40$

$\Sigma x = 198.91$ $\qquad \bar{x} = 4.97275$

$\Sigma x^2 = 993.5313$ $\qquad \frac{\Sigma x^2}{N} = 24.838283$

$s_x^2 = 0.110040$ $\qquad s_x = 0.331723$

General Summations from Frequency Tables

It seems desirable to emphasize at this point that, in our algebraic notation, Σ has been defined as indicating the sum of *all* the values directly following it. The symbol for the variable, for instance x, when placed after a summation sign is understood to indicate the individual values of x, the sum being made for all variate magnitudes.

The class centers in a frequency distribution really correspond to the variate magnitudes under an assumption that measurement was remade to the nearest value according to the scale of graduations given by those class centers. In the tables to be given immediately following in this

[1] *Vide* page 38.

chapter, these centers will be designated as x_c solely for purposes of distinction. This permits more convenient demonstration of the point that general summations made from frequency tables proceed by two steps. In securing Σx to calculate the mean, for instance, one first multiplies the class center by the frequency to get the sum of all x values in the class. Then one totals these products. Forgetting that multiplication is simply repeated addition of the same quantity, the reader might easily overlook the preliminary multiplication as being part of the general summation. If there are k classes in a frequency table, then Σx as applied to it is identical with $\overset{k}{\Sigma} fx_c$. We shall see that not only the first and second but also higher powers of x enter statistical computations, and summation of them called for by the general algebraic notation Σx^p reduces to $\overset{k}{\Sigma} fx_c^p$ in special frequency-table notation. The general rule may be stated as follows: If data are seriated into a frequency table and a sum over all N values for the variable is called for, each class center must be counted as many times as is indicated by the frequency of the class.

The Uncoded Frequency Table

Table 9 with its appended summary of derived values indicates the steps involved in securing the mean and standard deviation for data assembled in the form of a frequency table. The material is that already presented in histogram form as Fig. 8,[2] the seriation being that of panel (*D*). The first four columns of the table represent familiar steps. The last column, headed fx_c^2, is assembled by securing the products of the corresponding values in the x_c and fx_c columns, its sum being Σx^2 of the general algebraic notation. With a calculating machine there would not be any need to include this last column in the table as the products could be accumulated in a sum directly on the machine as each is secured.

Code Scales

Uniform grouping in a frequency table permits one to use code scales to speed up calculations. The standard deviation may be secured along with the mean on an x scale, the results then being transformed to the original scale. This transformation equation for converting s_x to s_x may be derived very simply as follows:

Since $\qquad x = a + b\mathrm{x}, \quad \text{and} \quad \bar{x} = a + b\bar{\mathrm{x}},$

then $\qquad x - \bar{x} = b(\mathrm{x} - \bar{\mathrm{x}}).$

[2] *Vide* page 31.

TABLE 9

CALCULATION OF MEAN AND STANDARD DEVIATION FOR WEIGHT (IN POUNDS) AT BIRTH OF 402 INFANTS

Class range	x_c	f	fx_c	fx_c^2
5.0– 5.49	5.25	7	36.75	192.9375
5.5– 5.99	5.75	28	161.00	925.7500
6.0– 6.49	6.25	56	350.00	2,187.5000
6.5– 6.99	6.75	72	486.00	3,280.5000
7.0– 7.49	7.25	86	623.50	4,520.3750
7.5– 7.99	7.75	65	503.75	3,904.0625
8.0– 8.49	8.25	42	346.50	2,858.6250
8.5– 8.99	8.75	22	192.50	1,684.3750
9.0– 9.49	9.25	13	120.25	1,112.3125
9.5– 9.99	9.75	7	68.25	665.4375
10.0–10.49	10.25	3	30.75	315.1875
10.5–10.99	10.75	1	10.75	115.5625
		402	2930.00	21,762.6250

$N = 402$

$\Sigma x = 2930.00.$ $\qquad \bar{x} = 7.288557.$

$\Sigma x^2 = 21762.6250.$ $\qquad \dfrac{\Sigma x^2}{N} = 54.125883.$

$s_x = 1.012820.$ $\qquad s_x = 1.006390.$

Now
$$s_x^2 = \frac{\Sigma(x - \bar{x})^2}{N}$$
$$= \frac{\Sigma[b(\mathrm{x} - \bar{\mathrm{x}})]^2}{N}$$
$$= b^2 \frac{\Sigma(\mathrm{x} - \bar{\mathrm{x}})^2}{N}$$
$$= b^2 s_{\mathrm{x}}^2.$$

Therefore,
$$s_x = b s_{\mathrm{x}}. \tag{5}$$

The coding constant a does not enter this transformation equation. Subtraction of a constant quantity from a set of variates will leave the residuals with the same variability as characterized the original measurements.

The full calculations for the mean and standard deviation of the series used in Table 9 are given through use of coding in Table 10. The sim-

TABLE 10

CALCULATIONS AS FOR TABLE 9 BUT USING CODING

x_c	f	x_c	fx_c	fx_c^2
5.25	7	0	0	0
5.75	28	1	28	28
6.25	56	2	112	224
6.75	72	3	216	648
7.25	86	4	344	1,376
7.75	65	5	325	1,625
8.25	42	6	252	1,512
8.75	22	7	154	1,078
9.25	13	8	104	832
9.75	7	9	63	567
10.25	3	10	30	300
10.75	1	11	11	121
	402		1,639	8,311

$N = 402.$ $a = 5.25.$ $b = 0.5.$

$\Sigma x = 1639.$ $\bar{x} = 4.077114.$

$\Sigma x^2 = 8311.$ $\frac{\Sigma x^2}{N} = 20.674129.$ $\bar{x} = a + b\bar{x} = 7.288557.$

$s_x^2 = 4.051270.$ $s_x = 2.012777.$ $s_x = bs_x = 1.006389.$

plification of the calculations by transformation of scale may again be attested by the ease with which the work in Table 10 may be verified by mental arithmetic. The final results are of course identical (within last-place modification errors) with those secured in Table 9.

"LEAST SQUARES"

In accepting the mean as the typical value about which to measure variation, and then establishing the standard deviation as an appropriate measure of the variation, we have associated two descriptive statistics that are held closely together by a natural mathematical bond which we are now in a position to reveal. Let us consider the change in the "mean square deviation" about any point a as the latter moves along the scale of measurement. Instead of subtracting $\bar{x}$ from every value of x, let us subtract some other constant, a. Then the mean square deviation about a will be

$$\text{M.S.D.} = \frac{\Sigma(x-a)^2}{N}$$
$$= \frac{\Sigma x^2}{N} - 2a\bar{x} + a^2$$
$$= \frac{\Sigma x^2}{N} - \bar{x}^2 + \bar{x}^2 - 2a\bar{x} + a^2$$
$$= s_x^2 + (\bar{x} - a)^2.$$

It is immaterial whether a is greater or less than $\bar{x}$ as far as the sign of $(\bar{x} - a)^2$ is concerned; this sign will always be positive. Therefore the mean square deviation will always be greater than the squared standard deviation unless a is made equal to $\bar{x}$. From this it follows directly that the mean of any series is the value from which the variates as a whole deviate the least, provided that variation is measured by squaring the deviations. The standard deviation measures the variation which has been so minimized. Mathematicians express this property of the mean by saying that it is the "least squares solution" of finding a value most closely approximating any series of magnitudes.

This affinity of mean and standard deviation may naturally raise the question: If the deviations had been taken without regard to sign instead of squaring them, would the mean still remain as the point about which variation so measured would be a minimum? Let the mean deviation be defined in general form as follows:

$$\text{M.D.} = \frac{\Sigma|x-a|}{N},$$

wherein the vertical rules indicate that the sign of the quantity enclosed by them is to be ignored. For what value of a is this mean deviation a minimum? The complexity of the derivation necessitates that this question be answered here without proof. The answer is quite certain, however, that the variates deviate least from their *median* value if first-power deviations without signs are used as the measure of variation. It seems very reasonable then that the "average deviation" should be taken about the median rather than about the mean; the median and this mean deviation have the same affinity as the mean and standard deviation.

In view of these facts and the probably greater simplicity of comprehension of the median and mean deviation, it is very reasonable to ask why statisticians work so generally in terms of the mean and standard deviation. The explanation arises from precisely the same type of reasoning as led to the question. We are in search of simplicity. "Least

squares" solutions, or the minimizing of squared deviations to reach representative values, may be made with facility, whereas the minimizing of absolute deviations becomes so involved that the problems can rarely be solved that way. Now problems of defining a moving representative value prove multitudinous in statistical analysis when one passes to the mutual relationships of two or more variables, and therefore consideration of the simplicity with which solution of such problems *in general* may be reached assumes paramount importance. If there is no difficulty in accepting the mean as a representative value, then its logical associate in measuring variation should likewise be acceptable. The use of the standard deviation as a measure of variation may at first offer the student some tasks in comprehension, but it will pave the way to progress where the average deviation would lead into a blind alley.

INDEPENDENT DEVIATIONS AND VARIANCE

During our preliminary discussion of the variation shown by N variates, attention was drawn to the point that fundamentally only $N - 1$ independent differences exist between variates. Of the $\frac{N}{2}(N-1)$ possible individual differences, $N - 1$ arise through comparing a selected variate with the others. The remainder arise from comparing the others with one another, all these secondary differences being derivable from what we may call the $N - 1$ primary deviations. Take a simple series of 3 variates, x_1, x_2, and x_3. The primary differences may be given as $x_2 - x_1$ and $x_3 - x_1$. The only possible secondary difference here is $x_2 - x_3$, which is just $[(x_2 - x_1) - (x_3 - x_1)]$, or the difference between the primary differences. This may be extended to a series of any size. No information about variation is gained by considering the secondary differences.

When deviations are taken from the mean instead of from some selected variate, quite obviously N differences (in this particular case we have called them deviates) arise. Has one additional element of information about variation been gained? Certainly not! Each deviate is to some extent smaller than it would be if we demanded *independent* deviations. Each deviate is, to a small degree, a comparison of a variate with itself, for $\bar{x}$ is the sum of the Nth parts of all variates. The extra deviate simply makes up for this. Fundamentally there are only $N - 1$ independent deviations. One might therefore argue that the sum of their squares should be divided by $N - 1$ instead of by N. This slight change in the divisor may be considered as inconsequential *unless N is small* and one is desirous of estimating the variation in the supply directly from the

sample. The adjustment plays a most important role in the "small-sample" techniques, to the development of which Fisher and others have contributed so notably. Detailed consideration of these techniques seems to the writer to belong as a sequel to an introductory discussion of statistical reasoning, and development of them will be reserved for another volume. The reader, however, is now in a position to understand a variant of the formula just given for the standard deviation. It is the square root of the *variance*, v, where

$$v_x = \frac{\Sigma(x - \bar{x})^2}{N - 1}. \tag{6}$$

The square root of v should *not* be called the standard deviation of the sample, but rather the maximum likelihood estimate (M.L.E.) of the standard deviation σ, of the supply. It may be derived very simply from s_x by means of a correction factor.

$$\text{M.L.E.}_\sigma = s_x \sqrt{\frac{N}{N - 1}}. \tag{7}$$

The reader may derive this factor for himself very readily from equations (2) and (6).

THE QUARTILES

Description of frequency distributions in terms of their "probability points" has been initiated in the discussion of the median as a central value. Its representativeness depends on its "middlemost" character. There is a very simple extension of this idea that permits of the portrayal of variation in the same terms.

Half of the variates in any series deviate in the positive sense from the median value, the other half being negative deviations. One may choose to establish the "middlemost deviation" in each direction. Such values taken with the median itself clearly divide the whole ranked series into 4 equal frequency groups. They are therefore called the *quartile* values, the lower being known as the first quartile (Q_1), the median forming the second, and the upper value the third quartile (Q_3). The value given by one-half of the difference between the first and third quartiles is used on occasion as a description of the variation present; it is usually called the *semi-interquartile range* and is symbolized as Q. Then

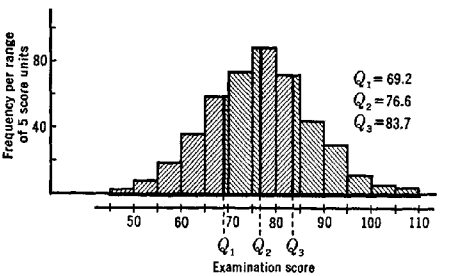
$$Q = \frac{Q_3 - Q_1}{2}. \tag{8}$$

It is worthy of note that, like the median, Q is not affected by extreme variates. It is a measure of mediocrity in variation, and is based on frequency of variates rather than individual deviation magnitudes.

When working from a frequency table rather than ranked variates, determination of the quartile values usually involves some calculation because the points must be assigned within class ranges. To clarify these matters an illustration of procedure is given in Table 11, utilizing the data on examination scores graphically portrayed in Fig. 2. The assumption underlying selection of an appropriate point within a class range is that the frequency of the class is uniformly distributed over the range. Attractive because of its simplicity, this assumption may be replaced, if desired, by others of more elaborate nature giving better approximation to frequency-curve form. We shall use the simple one herein. A "cumulative frequency" column is formed in which the frequency up to the boundary line between classes is given. The class range within which each quartile must fall then becomes apparent by inspection. That proportion of the scale covered by the class is taken which, under the assumption, yields the necessary frequency. Referring to Table 11, the first quartile will fall at the value which will cut off the lower 112.25 out of 449 units of frequency. Sixty-three units exist below a score of 65, and 122 below a score of 70. To 65 must be added the proportion $\frac{112.25 - 63}{122 - 63}$ of the class range immediately above 65. The evaluation of all quartile values is appended to the table. Figure 12 shows the quartiles plotted on the histogram, each shaded area comprising one-quarter of the whole, the dividing ordinates rising at the quartile values.

FIGURE 12

ILLUSTRATING THE QUARTILE VALUES IN A SET OF 449 EXAMINATION SCORES



* Data and calculations in Table 11, page 63.

TABLE 11

CUMULATIVE FREQUENCY AND THE CALCULATION OF QUARTILES
(Scores of 449 students in an examination*.)

Class range	Frequency f	Cumulative † frequency Σf
45- 49.9	2	2
50- 54.9	7	9
55- 59.9	18	27
60- 64.9	36	63
65- 69.9	59	122
70- 74.9	74	196
75- 79.9	88	284
80- 84.9	72	356
85- 89.9	44	400
90- 94.9	30	430
95- 99.9	11	441
100-104.9	5	446
105-109.9	3	449

$\frac{N}{4} = 112.25.$ $\quad Q_1 = 65 + \frac{49.25}{59} \times 5 = 65 + 4.2 = 69.2.$

$\frac{N}{2} = 224.5.$ $\quad Q_2 = 75 + \frac{28.5}{88} \times 5 = 75 + 1.6 = 76.6.$

$\frac{3N}{4} = 336.75.$ $\quad Q_3 = 80 + \frac{52.75}{72} \times 5 = 80 + 3.7 = 83.7.$

* Data by courtesy of the Committee on Educational Research, University of Minnesota.
† Cumulative frequency up to boundary between classes.

This same principle may be extended to the fractionation of the total frequency into any systematic scheme of parts desired. Among these, *deciles* and *percentiles* are not infrequently used in discussions of educational data. The reader may readily gain some experience with these finer divisions and the problems they present by first applying the above principles of procedure to the data of Table 11.

RELATIVE VARIATION

For the comparison of the variabilities existent in different series, it is sometimes necessary to express variation on a relative scale. For example, the standard deviation of finger length (left middle finger) for the series of Fig. 1 is 0.5479 mm. For the same group of individuals, the standard deviations for head length and stature were found to be 6.046 mm. and 2.5410 inches, respectively. Direct comparison of these standard deviations is not very informative because of the difference in general level of magnitude of these dimensions. Moreover, one is expressed in inches and the others in millimeters. If, however, the standard deviations are expressed as percentages of the respective means, yielding 4.74 per cent, 3.15 per cent, and 3.88 per cent, respectively, it is seen immediately that finger length is relatively the most variable, and head length relatively the least variable, of the three. This relative measure of deviation is known as the *coefficient of variation*:

$$\text{C.V.} = 100\frac{s_x}{\bar{x}}. \tag{9}$$

A word of warning is warranted with reference to the use of this measure. In its interpretation, caution must always be exercised not to neglect the fact that its magnitude is just as much a function of the mean as it is of the standard deviation. Variabilities measured on fully comparable scales should be compared directly in terms of the standard deviations as far as is practical. If a relative measure is necessary in order to circumvent the use of different scales or to allow for incompatibility of means, care should be exercised to assure oneself that it is wholly logical to consider the variation in proportion to the mean value as a basis of comparison. The use of the coefficient of variation may give valuable information on questions such as the following: Is human stature more variable in the adult than in the new-born infant? Is the error of the chemical determination of ash in foods relatively greater than that for protein in the same materials? It would be quite foolish to compare the error of a measurement process with the variability of individuals so measured, by means of the coefficient of variation.

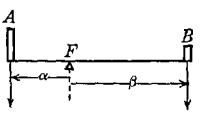
TYPE OF VARIATION

The foregoing discussion of variation has been strictly confined to the question of *amount* of variation. Questions of skewness, kurtosis, *et cetera*, have been intentionally divorced from the discussion, for they are concerned with *type* of variation. In considerations of the amount of variation, the signs of the differences between variates are irrelevant; one is concerned solely with consistency of measurements, like the scattering of bullet holes about their focal point on a target. We shall now pass to a general scheme of measuring the characteristic features of frequency distributions in which $\bar{x}$ and s_x form the first elements, measures of skewness and kurtosis following directly in the scheme.

CHAPTER 5

MOMENTS AND DISTRIBUTION CHARACTERISTICS

The student of mechanics very early becomes familiar with the concept of "the *moment* of a force about a point," or the importance of the force in producing (or tending to produce) motion about an axis. In games of "seesaw," a child learning to balance another on a beam, which itself is balanced about a "fulcrum," discovers that the effectiveness of its weight on the beam is directly proportional to its distance from the fulcrum. A sketch of such a system in equilibrium is given below for reference. When equilibrium is established it is found that the product of a weight A and its distance α from the fulcrum F is equal to the similar product for B. These products are termed "moments" in mechanics. The two weights A and B exert their force in the same downward direction. The distances α and β are measured in opposite senses and therefore are appropriately given opposite signs. The two moments then are $-A\alpha$ and $+B\beta$, the sum being zero when equilibrium prevails.



The effectiveness of a weight on a horizontal beam[1] with respect to its tendency to produce motion may similarly be measured with respect to any point in the beam. The fulcrum may be moved to any point, each such point of reference being known as an *origin* of the calculated moment. If there are k objects on a weightless rod which is scaled in equal increments from some point (for example, one end, although that is immaterial), then the total moment of the system about any fixed origin a in the beam is given by the expression $\Sigma f(x - a)$, where f is each weight, x is its position point on the scale, and summation is made for all k products.

Let x be the scale of a variable, of which a sample of total frequency N has been seriated into k classes, f designating the frequency of any class. A histogram drawn for this seriation may be visualized as a series of rectangles on a weightless beam, the weight of each rectangle being proportional to its area, that is, the class frequency in each case. If a is

[1] The beam is theoretically assumed to be without weight of its own.

the point in the rod about which the system would be in balance if a fulcrum were placed there, then obviously a lies somewhere between the lowest class value, x_1, and the highest class value, x_k. The value of a may be determined very simply, since the condition of equilibrium demands that the total moment shall be zero. The moment of each class is $f(x - a)$, and by definition

$$\Sigma f(x - a) = 0.$$

Since the only variable quantities here are f and x, the expression may be expanded to the form

$$\Sigma fx - a\Sigma f = 0.$$

But $\Sigma f = N,$

and $a\Sigma f = Na.$

Therefore $\Sigma fx = Na,$

or $$a = \frac{\Sigma fx}{N} = \bar{x}.$$

Thus the mean of any series is the balancing point with respect to weight of the observations on the scale of measurement.

The general characters of frequency distributions which have already been noted, namely,

(1) total frequency,
(2) balancing point,
(3) amount of variation,
(4) skewness, and
(5) kurtosis,

may all be defined very readily in terms of a simple series of extensions of this moment concept. These extensions emanate from the basic step of considering not only the distance factor, $x - a$, but also the series of powers from zero through the fourth of this factor. Let us analyze the expression for the mean squared deviate, from which our standard measure of variation is derived. Using special frequency-table notation (x_c for class centers),

$$m_2 = \frac{\Sigma f(x_c - \bar{x})^2}{N}.$$

This is simply the mean *second* moment about the origin $\bar{x}$. The order of the moment, designated by the qualifying adjective, is the power to which the distance factor or deviation is raised. If each value of x is

considered individually the equation reduces to the general algebraic form

$$m_2 = \frac{\Sigma(x - \bar{x})^2}{N},$$

f being reduced to the constant 1 and therefore omitted.

One may conveniently generalize the terminology for statistical moments according to the following scheme:

(1) The nth moment of an *individual* x about an origin a on its scale of measurement is $(x - a)^n$.

(2) The *total* nth moment of a *series* about the origin a is $\Sigma(x - a)^n$.

(3) The *mean* nth moment of a *series* of N individuals about an origin a is $\frac{\Sigma(x - a)^n}{N}$.

The statistician in his studies of variables is concerned with average values derived from series of individuals, that is, with single figure representations of mass characteristics. The moment of an individual and the total moment of a series are but steps, as it were, to an average moment of some sort. Also he is very largely concerned for reasons of convenience with two particular origins, the mean value of the series and the scale origin of zero. In the *theoretical* development of statistical concepts it is most convenient to refer measurements to the mean as origin, and as a consequence a special designation in word and symbolism has arisen for the average moments about the mean as origin. These are known as the *moment coefficients*, and the symbols used for them are m with definitive subscripts. Hence the previous use of m_2, meaning the second-moment coefficient, as the symbolic designation for the mean squared deviate. We shall later use m' as the symbol for a mean moment taken about zero as origin. In accordance with this notation,

$$m_1' = \frac{\Sigma x}{N} = \bar{x},$$

$$m_2' = \frac{\Sigma x^2}{N},$$

and so forth.

It is interesting, though quite incidental, to note that one total moment is used as such, namely, the total zeroth moment. This is the total frequency of any series, designated herein as N. It is a derived property of numbers that *any* quantity when raised to the zeroth power

is unity. Thus, $(x - a)^0$ equals unity, and therefore $\Sigma(x - a)^0$ equals N; the frequency of a series is a total moment, and all mean moments are therefore ratios of total moments.

One may hope that the utility of the lower moments for describing properties of frequency distributions will by now have appealed to the reader. The zeroth defines frequency, the first defines the most broadly useful measure of type, and the second provides the standard measure of variation. When it is realized that the characteristics of skewness and kurtosis in frequency distributions may likewise be defined very simply in terms of the third- and fourth-moment coefficients respectively, some appreciation of the power of analysis by means of moments will have been achieved. Before proceeding with demonstration of this, however, perhaps one might be pardoned for a digression into the historical basis for the use of the term "moment."

It was while encumbent of the professorship of applied mathematics and mechanics at University College, London, with students of engineering chiefly forming his classes, that Karl Pearson laid many of the foundations of modern statistical analysis. Challenged by visions implanted by Francis Galton's masterful book "Natural Inheritance" (1889), and immensely stimulated by Weldon (professor of zoology at University College) and Galton to blaze trails on the frontier of application of mathematical reasoning to the resolution of *biological* problems, Pearson faced the task of objectively defining more precisely and more completely the characteristics of frequency distributions. In extension of the mechanical concept of moments he perceived possibilities for the development of powerful statistical tools for this purpose, possibilities which he soon developed into realities. The zeroth, first, and second moments were already in use though not recognized as such. Let us return now to a consideration of the utility of the third and fourth moments.

SKEWNESS

In perfectly symmetrical distributions it is easy to perceive that all the odd-moment coefficients must be zero. The odd powers of $x - \bar{x}$ retain the sign of the deviate itself, and, each positive deviation being exactly balanced in frequency by that of the deviation of opposite sign, the sum must reduce to zero. Perfect symmetry in a random sample from a symmetrical supply of variates would, of course, not be expected except as a coincidence. However, the third-moment coefficient for such random samples in general will tend toward zero and differ from that value only within the relatively small margin of sampling errors.

On the other hand, if the distribution of frequency is skew this condition will not apply. Equal magnitudes of positive and negative deviates will have systematically differing frequencies, and the cubing of the deviates, giving rapidly increasing importance in the sum to each deviate as its magnitude increases, will reflect the skewness in the more or less marked departure of the third-moment coefficient from zero. The effect is that of making the "tail wag the dog." The long tail of the distribution will contribute more to the total than the deviates of opposed sign can offset, and so the total third moment and m_3 will assume the sign of the direction of the long tail. Also, the longer the tail (relatively), or the greater the skewness, the larger will m_3 become. The same feature will prevail with every higher *odd* moment; the higher the moment the greater the magnitude of the sum or coefficient. It is simpler, of course, to use the lowest power that will achieve the purpose satisfactorily, and so m_3 forms the basis of measuring skewness.

It is perhaps helpful to consider here a concrete example of skewness. The data in Table 12 represent an arbitrarily defined frequency distribution of 100 individuals on a simple scale. It has been devised as a perfectly symmetrical distribution. By shifting two individuals in this distribution, one from class $x = 2$ into class $x = 0$, and one from class $x = 3$ into class $x = 5$, perceptible skewness has been introduced with-

TABLE 12

ILLUSTRATING THE EFFECT OF SKEWNESS ON THE THIRD MOMENT

Class center x	Frequency f		Deviate $x-\bar{x}$	Moments			
				first $f(x-\bar{x})$		third $f(x-\bar{x})^3$	
0		*1*	−4		*− 4*		*−64*
1	2	*2*	−3	− 6	*− 6*	−54	*−54*
2	9	*8*	−2	−18	*−16*	−72	*−64*
3	23	*22*	−1	−23	*−22*	−23	*−22*
4	32	*32*	0	0	*0*	0	*0*
5	23	*24*	+1	+23	*+24*	+23	*+24*
6	9	*9*	+2	+18	*+18*	+72	*+72*
7	2	*2*	+3	+ 6	*+ 6*	+54	*+54*
Totals	100	100		0	*0*	0	*−54*
Moment coefficients				0	0	0	*−0.54*

out altering the mean value, $\bar{x} = 4$. The frequencies and calculations for the skew series are given in italics in juxtaposition with those of the symmetrical distribution. The effect of the shift of two individuals on $\Sigma(x - \bar{x})^3$ and m_3 may now be studied numerically. For the symmetrical distribution, $\Sigma(x - \bar{x})^3$ and m_3 are obviously zero. For the skew distribution, however, they are negative. This is purely a result of raising the deviates to a higher odd power than the first, wherein they balanced regardless of form of distribution.

KURTOSIS

When the even moments are considered, it will be recognized that, although the effect of weighting larger deviates through using higher powers persists, the class moments now all have the same sign. Thus the even moments cannot reflect skewness; the result is the same as if the negative half of the distribution of deviates were folded over and added to the positive half. The effect of higher even powers is to accentuate the relative importance of the extreme deviations regardless of their signs in the first power. For this reason the standard deviation is always greater than the average deviation.[2] If the fourth power of the deviates is used, still greater emphasis will be given in the sum to the presence of relatively long tails in a distribution, regardless of skewness. Here, then, we are dealing with relative concentration of the variates about the mean. Two distributions having the same standard deviation will have different values of m_4 if one has longer tails than the other, the higher m_4 going with the "longer-tailed" distribution. To get such longer tails and keep the standard deviation the same, it is necessary to move frequency from the medium deviation group (which we may call the "shoulders" of the distribution), placing some individuals nearer the mean and some farther away. Thus "peakedness" at the center is increased along with tail extension when m_4 is increased without altering the *amount* of variation. It was because of this that the term *kurtosis* (implying sharpness of curvature at the center) was introduced by Karl Pearson for this feature of relative clustering about the mean.

The use of powers of deviates has introduced a method of quantitatively describing certain readily recognizable characteristics of frequency distributions. Each mean moment secured is, however, a count of units which are themselves the appropriate powers of the original unit of

[2] For reasonably normal distributions the standard deviation will be found to be approximately 25 per cent greater than the average deviation. For the normal curve itself the relationship is

S. D. = 1.2533 A. D.

measurement of the variable.[3] Skewness and kurtosis, on the other hand, are concepts that are basically independent of any scale of measurement of a variable, and description of them should logically proceed on a scale of pure numbers. Only in this manner may degrees of both skewness and kurtosis be quantitatively described in terms which are directly comparable from one distribution to another, regardless of the scales of measurement of those variables.

Elimination of the units of measurement from moment coefficients may readily be secured by taking suitable ratios of them. Since the immediate problem is to transform the m_3 and m_4 values to pure number scales, and m_2 is the lowest moment coefficient not having a fixed value, ratios involving m_2 as the simplest reference unit would appear to offer possibilities. Karl Pearson suggested two such ratios which have proved very serviceable and are widely used. Employing Greek letters where we use the English forms, he called them the "beta coefficients." In our symbolism they may be written

$$b_1 = \frac{m_3^2}{m_2^3}, \text{ an index of skewness,} \tag{1}$$

and

$$b_2 = \frac{m_4}{m_2^2}, \text{ an index of kurtosis.} \tag{2}$$

Both numerator and denominator of the ratio for b_1 involve sixth powers of the original unit of measurement, and therefore the coefficient is a pure number. Likewise, b_2 is a pure number derived from fourth-power units. The utility of such empirical ratios should be judged in terms of how well in practice they serve the purpose of description for which they are offered. The answer arising from experience is that basically they serve very well indeed. Certain slight modifications, aiming to introduce direction of departure from normal curve values, are preferred by some. These latter indices are designated as the "gamma coefficients" or "g statistics":

$$g_1 = \frac{m_3}{s^3}, \text{ an index of skewness,} \tag{3}$$

and

$$g_2 = \frac{m_4}{s^4} - 3, \text{ an index of kurtosis.} \tag{4}$$

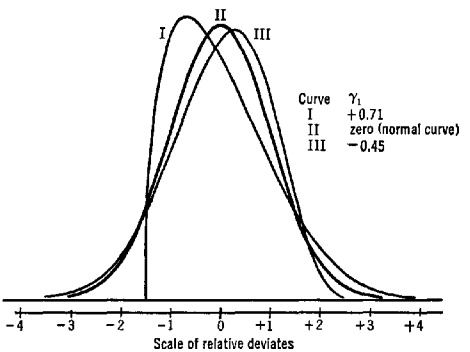
[3] The dimension "12 inches" when squared yields "144 *square* inches," and so on for higher powers. Extracting the square root of m_2 to get the standard deviation had, as its objective, a return to the original scale of measurement.

Whereas the value of b_1 is zero for symmetrical distributions (since $m_3 = 0$), it is positive for all other curves because m_3 was squared to secure it. g_1, on the other hand, preserves the sign appropriate to the direction of skewness. Obviously, squaring g_1 gives b_1.

Figure 13 presents three curves having different degrees of skewness but otherwise of like character. The b and g coefficients are given in proximity to these curves, the middle curve being the symmetrical normal curve.

FIGURE 13

γ_1 AS A MEASURE OF SKEWNESS



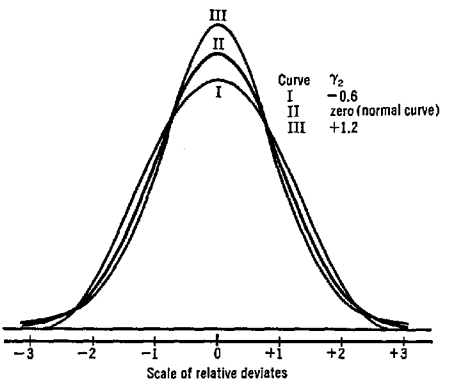
The fourth-moment coefficient, m_4, is always positive, and therefore b_2 does not automatically provide a central reference value of zero. Since the normal curve is a logical central type from which departure may be measured with respect to kurtosis as well as skewness, and it has the simple whole-number value of 3 for its b_2 coefficient, kurtosis may be measured directly from this origin.

$$g_2 = b_2 - 3.$$

Thus g_2 introduces the idea of negative and positive kurtosis with respect to the normal curve as a standard of reference, positive and negative values of g_2 automatically specifying greater or less central clustering or "peakedness" than the normal curve possesses. Figure 14 presents three superimposed symmetrical I-shaped curves differing in kurtosis but otherwise similar. All curves have the same mean, standard deviation, and area, the gamma coefficients changing directly with the degree of central clustering.

FIGURE 14

γ_2 AS A MEASURE OF KURTOSIS



CURVE FITTING

The subject of fitting frequency curves to observed distributions is a very interesting one from the theoretical point of view. It has rather small practical value, however, in statistical investigations which are not of a specialized nature demanding such work. Present comment on the subject will therefore be confined to indication of the mode of procedure in what is probably the most widely used method of fitting frequency curves, that given by the Pearsonian system.

When it is desired to fit a frequency curve to an observed frequency distribution a decision must be made as to the particular respects in which the curve *must* fit. That is, a curve is fitted to a histogram by selecting an equation for the curve which has certain properties identical with corresponding properties of the histogram. The larger the number of these identical values, the better the fit will be in general, but a curve may be "overfitted" in the sense that the curve itself reproduces the histogram too well, preserving rather than smoothing out its errors of sampling. Thus a compromise is always necessary; as few identities are made as possible to secure reproduction of the essential features without overemphasizing small details. The system of moments leads very nicely to the establishment of descriptions of distribution characteristics on a broad basis, and Pearson has amplified their usefulness by elaborating a scheme of curves covering a wide array of combinations of these moments to the fourth order.

Starting with a differential equation which yields the normal curve as one of its solutions, Pearson derived a series of frequency functions which appeared to be adaptable to nearly all observed types of frequency distribution. The two beta coefficients already defined, together with mean, standard deviation, and total frequency, are all the values needed from an observed distribution in order to select an appropriate curve from Pearson's series for graduation purposes. The technical discussion of procedure in fitting is quite beyond the scope of this book. The interested reader may be referred for such detail to Elderton[4] and to Pearson's summary of procedure[5] with most useful graphs for selecting the appropriate type of curve.

[4] W. Palin Elderton. Frequency curves and correlation. Cambridge University Press. 3d edition, 1938.

[5] Karl Pearson. Tables for statisticians and biometricians, Part II. Cambridge University Press. 1931.

CHAPTER 6

THE NORMAL CURVE

At the close of the second chapter the statement was made that biological variables for the most part do not deviate very markedly from the "normal curve" or "law of error" form. Both Quetelet and Galton, working with anthropometric variables, were deeply impressed by the way this curve seemed to graduate their data. The prominence they gave it probably contributed more than anything else to an early notion that this curve (which Galton often referred to as the "normal law") was the normal type for biological variables, whereas we now recognize it only as a central type in a general scheme of biological frequency distributions. The designation of "normal curve" was accepted by Pearson as a convenient way of avoiding the issue of choosing between the originator titles "Gaussian curve," "Laplacian curve," and "Gauss-Laplace curve," to which must now be added "DeMoivrian curve." In adhering to the term *normal curve* herein, one wishes to warn the reader against any implication that other types of curves are "abnormal" in the usual sense of the word.

While holding considerable interest as a graduating function that works remarkably well for many biological variables, the normal curve is far more important for its portrayal of the way in which means and other statistics vary through errors of random sampling. It commands so much interest that a somewhat more detailed analysis of its nature than has so far been presented seems appropriate at this point.

Mathematical equations defining the normal curve may all be reduced to the simple form

$$w = Ce^{-\frac{1}{2}k^2} = \frac{C}{e^{\frac{1}{2}k^2}}, \tag{1}$$

where w = the ordinate at any point x,
C = a constant of the curve,
e = the natural logarithm base, 2.7183 $\cdots$, and
k = the *relative deviate* of x from the mean of the curve.

Even this equation may look complex enough to the biologist on first inspection, but if he is willing to refresh himself a little on exponents he may analyze the form of the curve without difficulty.

First of all, let us consider the transformation from the variate scale, x, to the *relative deviate scale*, k. The largeness or smallness of any value x may be expressed on the universal scale of pure numbers by securing its relative deviate value. This is simply the *number of standard deviations* that x is away from the mean of the series, either in the positive or negative direction.

$$k = \frac{x - \bar{x}}{s_x}. \tag{2}$$

Both $x - \bar{x}$ and s_x are dimensions on the scale of measurement of x. If we divide one by the other the original units of measurements (inches, pounds, *et cetera*) vanish, and k as defined above becomes a pure number. This is a generalizing transformation of scale of great importance in statistics. It enables one to compare the largeness (or smallness) of a variate in its distribution with the same property for any other variate, whatever the variable may be. The transformation will be of great value to our development of the notion of correlation in the following chapter. At the present moment, however, it serves to place a standard scale under *all* normal distributions, permitting our study of the properties of the curve as such. Once the mean and standard deviation of a variable have been secured, the universal scale of relative deviates may be placed in juxtaposition with its distribution. This k scale will have its 0 value against $\bar{x}$, the $+1$ and -1 values respectively at the deviations of plus and minus one standard deviation s_x from $\bar{x}$, the scale extending without limit by such steps in both directions.

The natural base of logarithms, e, is of no consequence to us at the moment beyond recognition of the fact that it is a pure number greater than unity. Some knowledge is necessary, however, of the numbers which result when such a number is raised to powers. Since the zeroth power of any number is 1, and the first power of e is 2.7183, then any power of e between the zeroth and the first yields a number between 1 and 2.7183. Powers of e above the first yield numbers greater than 2.7183, of course, and all positive powers of e will therefore yield numbers greater than 1. We are now in a position to analyze the form of the normal curve as it is defined in equation (1).

ORDINATES OF THE CURVE

The ordinate of the normal curve will reach its *greatest* value C when $e^{\frac{1}{2}k^2}$ reaches its minimum value of unity. This minimum value occurs only when k is zero, corresponding to the mean value on the scale of any distribution of reference. Since k is squared in equation (1) the ordinate at any chosen negative value of k must equal the ordinate at the same

positive value of k and be less than the value C at the center. Therefore the distribution curve is of the symmetrical I form.

Those familiar with logarithms may readily calculate a series of ordinates and draw the curve. Arbitrarily making the mid-ordinate 100, one finds the ordinates at 1, 2, and 3 standard deviations each way to be 60.7, 13.5, and 1.1 units approximately. Larger values of k than 3 will give very small values of w, which, nevertheless, will be real; so the curve has a theoretically infinite range in both directions from the mean. This does not invalidate the curve as applied to biological data, for it may readily be seen that the probability of extreme deviates in normal distributions approaches infinitesimal magnitudes quite in accord with statistical experience of biological variables.

A plot of the normal frequency function scaled in terms of relative deviates is given in panel (*A*) of Fig. 15. Ordinates have been drawn at distances of 1, 2, and 3 standard deviations on either side of the mean. It will be noted immediately that the curve appears to merge into the base line at, or a trifle beyond, 3 standard deviations. At the points k equals $+3$ and -3, the height of the ordinate is only one-ninetieth part of that at the center, and the area in each tail so truncated is only a little over one-thousandth part of the whole area. For practical purposes the tangible range of the normal curve is accordingly often referred to as being roughly three standard deviations *on both sides of the mean.* As was noted above, the range is theoretically infinite, but the parts beyond ± 3 standard deviations are barely perceptible.

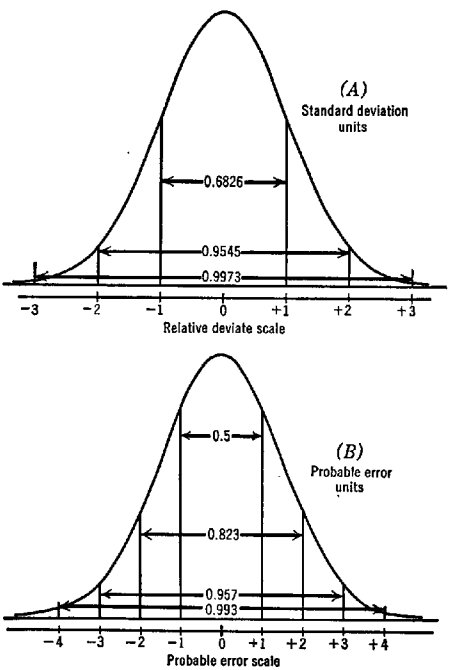
PROBABILITIES OF THE CURVE

In Chapter 2, emphasis was given to the point that, for continuous variables, frequency is properly portrayed by area. The frequency curve being the appropriate form of graduation for such variables, it follows at once that the area between any two ordinates, when expressed as a proportion of the entire area of the curve, defines the probability of occurrence of values between the two points on the scale at which those ordinates are drawn. This is a very important concept in statistics—indeed, the one upon which the final statistical verdict in many analyses rests. The normal curve plays so crucial a part in determining such verdicts that it is important for the student to acquaint himself somewhat with the probability zones of the curve. It is partly for this reason that the ordinates establishing such zones have been drawn in panel (*A*) of Fig. 15. By means of a planimeter or other suitable device the areas between the ordinates in the figure could be determined. If these were then expressed as proportions of the total area of the curve, it would be

found, for instance, that a little more than two-thirds of the total area lies between the ordinates erected at ± 1 standard deviation, and approximately 95 per cent lies between ± 2 standard deviations.

FIGURE 15

RELATIVE AREAS BETWEEN SYMMETRICALLY PLACED ORDINATES FOR THE NORMAL CURVE



These proportionate areas may be evaluated more exactly, indeed to any required degree of precision, by means of the integral calculus, and we shall rely now upon determinations so made. For the central zones indicated the values are given to four places of decimals. Note that 99.73 (or more precisely 99.73002$\cdots$) per cent of the area lies within ± 3 standard deviations; that is, the probability of a normally distributed variable having deviates exceeding ± 3 standard deviations is slightly less than 27 in 10,000.

The ordinates embracing the central 95 per cent of cases, or truncating the extreme 2.5 per cent of the total area in each direction, have come to play an important part in statistical interpretation. These ordinates arise at the k values of $\pm 1.9600\cdots$, or what is roughly called two standard deviations on either side of the mean. That is, in a normally distributed variable, there is only 1 chance in 20 that values will deviate from the mean by *more* than 1.96 standard deviations; the chance is actually 1 in 22 that values will deviate from the mean by more than two standard deviations.

The question may be raised: At what distance from the mean would the two ordinates embracing the central 50 per cent of the area be placed? Such ordinates will define the zone within and outside of which it is *equally probable* that individuals will occur. Those ordinates arise at the points k equals $\pm 0.6744898\cdots$. The distance of these ordinates from the mean was extensively used by astronomers in describing the amount of error made in measuring the stellar universe. Later adopted by the biologists as a measure of variation, the quantity $\pm 0.6745\ s_x$ has become commonly known as the *probable error*. The reader may note that the probable-error points correspond to the first and third quartile values.

Panel (B) of Fig. 15 shows a normal curve divided into zones by ordinates raised at ± 1, ± 2, and ± 3 probable errors. Since the factor 0.6745 differs only very slightly from two-thirds, the "5 per cent points" are approximated roughly by three probable error units, just as they are by two standard deviation units. These approximations are very useful to remember as simple guiding values in judging the importance of a deviation.

TABLES OF THE NORMAL CURVE

One's readily available information on probabilities in general for the normal curve needs to be much more detailed than these few skeleton values. A very frequently recurring problem in statistical analysis, especially when the variation is attributed to "errors of random sampling," is to ascertain with reasonable precision the probability that a specified value in a normal distribution will be exceeded. Occasionally, also, the normal curve ordinates are desired for plotting the curve. Accordingly, tables of these two functions are very common indeed. Such tables need to be in a generalized form so that they may be adapted to any normal distribution with its specified values of the mean and standard deviation. The most convenient arrangement is to give the ordinates and fractional areas for the curve in terms of k, the relative deviate, as a scale of measurement, assigning a total area of one unit to

the curve. Just as the use of relative deviates fixes the base scale, so the requirement that the area under the curve shall be unity fixes the vertical scale. Very full tables of ordinates and areas for the normal curve are so given in Part I of Pearson's "Tables for Statisticians and Biometricians," and those concerned with fitting the normal curve or securing probabilities with considerable precision are referred to that source.

In determining the probability that an event may be expected to occur through "errors of random sampling," one is often quite satisfied with three or four places of decimals in that probability. Appendix I to this book gives for such purposes a table of the needed relative areas to four places of decimals, by increments of one-hundredth in k. The ordinates are also given in juxtaposition to three places of decimals. The area or probability, P, given in Appendix I, is that in the tail of the curve beyond the point k, expressed as a proportion of the area of half the curve. This is the same probability as that of both tails combined in relation to the whole curve. Discussion of the reason for choosing this particular probability function is given in the Appendix.

SOME APPLICATIONS OF THE NORMAL CURVE TO OBSERVED DISTRIBUTIONS

Although it is interesting to fit the appropriate curve to any sufficiently large series of variates, it is not usually of any particular value to do so in practical analysis of data. Exercises in fitting a normal curve, therefore, become of academic interest rather than scientific importance. The situation is very different, of course, when one needs to consider the probabilities of occurrences obeying the normal law, the answers to these propositions being derived from the tables of the normal curve areas. Those engaged in precise measurement may well follow with very keen interest the probability of the occurrence of a chance error of any given magnitude. If the errors are unbiased and normally distributed, knowledge of the mean and standard deviation of a sufficiently large series of repetitions of the measurement will give the key at once to those probabilities. In this connection the reader may choose to apply himself to such questions with respect to the width of the spectral-line data given in Fig. 1[1]. The mean and standard deviation of width of line for the measurements are 178.9 and 3.583 microns, respectively.

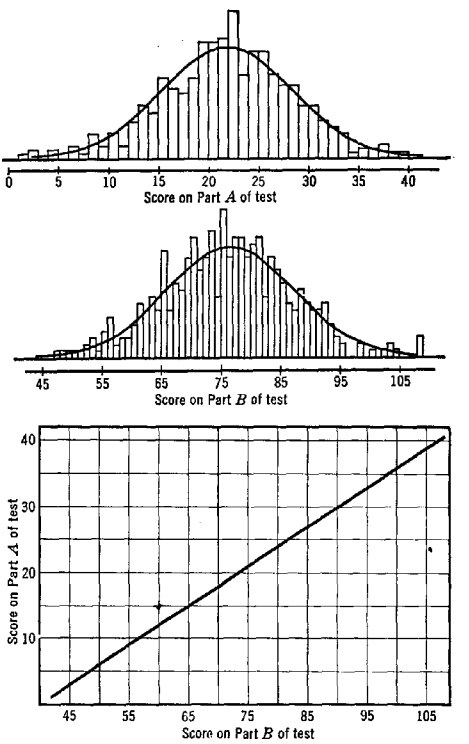
A problem of not infrequent occurrence in educational work is that of converting scores on aptitude or intelligence tests to their equivalent values on some more or less standard score scale, or to their equivalent

[1] *Vide* page 13.

values on the scale of any other specified test of like purpose. We may illustrate the underlying reasoning appropriate to such transfers in terms of application of the normal curve. The distribution of scores on

FIGURE 16

DETERMINATION OF EQUIVALENT SCORES IN TWO EXAMINATIONS BY EQUATING THE RELATIVE DEVIATES



Part B (current affairs) of a "reflective thinking" examination given to 449 students has been used in Fig. 2b[2] as an illustration of the commonly occurring normal distribution of measurements of specific intellectual achievements. They are reproduced in panel (B) of Fig. 16, wherein the histogram follows the original marks without grouping. The dis-

[2] *Vide* page 14.

tribution of scores for the same students on Part A (science) of the examination is given in like form as panel (A) of Fig. 16. In both, a normal curve graduates the data satisfactorily. The actual scores on Part A varied from 1 to 39; those on Part B covered the much wider range from 47 to 108. For each score on Part A, what is the comparable score on Part B as far as performance of the group as a whole is concerned?

Equating the possible ranges of marks, 0 to 49, and 0 to 123, respectively, would not answer this question, for that would assume the two examinations to be of equal difficulty. Equating the observed ranges of scores would also fail to answer the question, for it would ignore all marks but the extreme ones—an obvious blunder. One may solve the problem appropriately only by equating those scores which are of equivalent probability of occurrence as indicated by the graduating curves. Since both curves are of the normal type, this is a simple problem; one needs only to find the scores for each test which have the same relative deviate values. The two distributions have been drawn in Fig. 16 with their means and standard deviations in correspondence. This gives to the corresponding points on the two scales equivalent probability according to the graduating curves.

Panels (A) and (B) of Fig. 16 have been prepared solely to demonstrate the underlying principle. They are by no means necessary for solution of the problem. A simple graph with one score scale along the axis of abscissas and the other along the axis of ordinates might have been prepared. A straight line drawn across the surface and passing through the points $(\bar{x}, \bar{y})$ and $(\bar{x} + s_x, \bar{y} + s_y)$ would establish the equivalent scores immediately. Such a graph forms panel (C) of Fig. 16, from which the equivalent scores may be read off with ease.

It may be noted that the problem of finding scores of equivalent probability of occurrence is simplified in the above illustration because both sets of scores are normally distributed. The same procedure might also be followed when both distributions have the same basic form as evidenced by their "gamma" coefficients. However, distributions of essentially different form would necessitate rather involved calculation to find the equivalent probability scores.

Consideration of the use of the normal curve in the interpretation of the significance of differences between means and other descriptive statistics must be deferred until discussion of the "errors of random sampling" is undertaken in a later chapter. Since acquaintance with the correlation coefficient will prove most important to adequate development of the principles involved in such tests, attention will first be directed to the description of associated variation.

CHAPTER 7

BIVARIATE DISTRIBUTION AND THE COEFFICIENT OF CORRELATION

Study of the variation in characters considered singly leads naturally to inquiries concerning the tendency of related characters to show associated variation. Baking quality of flour is a variable characteristic of that product of the milling of wheat. Likewise the chemical composition of wheat varies from sample to sample. Is the former related in any manner to the latter? To what extent does the yield of sugar from a field of beets depend upon the rainfall or the quantity of fertilizer employed? Is the intelligence of the child more highly related to that of the mother than to that of the father? Does environment influence the error of personal judgment? A veritable host of such questions must have suggested themselves at one time or another to the imaginative student of biology.

Francis Galton, however, was the first to give a practical solution to the problem in the form of a readily determinable and widely comparable numerical measure of the intensity of association between variations. He introduced to quantitative biology a new concept that has been most useful to a large body of statistical procedure. It has been revolutionary in its effect of displacing the idea of complete causation by one of incomplete association, the intensity of which may be measured with precision. It is a great step forward in science to proceed from the measurement of things to the quantitative determination of the degree of interdependence between things. Galton's contribution to science in this regard has had consequences reaching far beyond his specific problem of human inheritance. Among statisticians the importance of the concept itself is too often lost sight of in disputations concerning the merits of various statistical solutions.

The problem of measuring associated variation is again one of describing the character of a frequency distribution. This time, however, the distribution must be extended in its spatial formation, since one dimension will be required for each variable considered. For the interrelation of two variables, the frequency distribution may be designated as *bivariate* by way of distinction from the *univariate* systems so far considered.

PORTRAYAL OF THE CORRELATION SURFACE

The graphical representation of a bivariate frequency distribution does not present any difficulty. Two dimensions of space are required for the scales of measurement of the two variables, and these may be chosen conveniently as the axes of ordinates and abscissas of any cross-section paper. Each point on the surface so scaled will correspond to a pair of magnitudes of the variables. The task remains to find a suitable form of representation for the frequency of occurrence of these paired values. Undoubtedly, understanding of the correlation surface would be considerably aided if the third dimension perpendicular to the surface could readily be used for this purpose, but the preparation of such a solid is sufficiently difficult to forbid the general practice. Quite obviously, it is necessary in most cases to confine the representation to the two dimensions of a surface.

Two methods for the representation of frequency under such circumstances without demanding another dimension of space are quite familiar to all. First, a dot may be used for each pair of associated variates. The optical effect of the clustering of the dots in such a *scatter diagram* suggests density of population at once, and imparts to the device a decided superiority for graphical purposes. Second, the frequency of individuals occurring within prescribed class ranges may be given in numerical form, thus providing a *correlation table.* This is a decided aid in computational work when the number of paired variates is large.

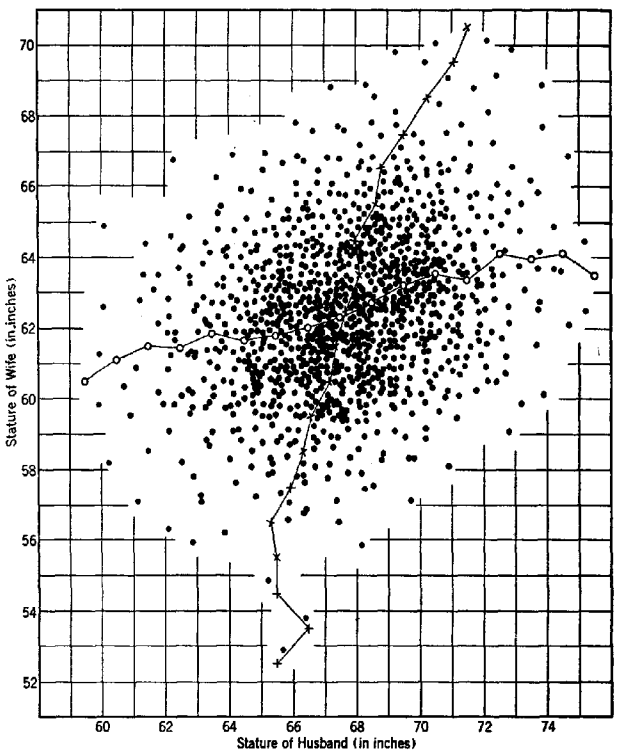
One may consider, by way of illustration, the distribution of frequency in the assortative mating of man for stature. Do tall men seek the companionship of tall women, or is there a law of opposites that makes "brunettes and blondes rush irresistably together"? The existence of such a law is not merely an assumption of Percy Public; it was taught not many years ago in the social sciences. That which is unusual claims interest, and the commonplace most frequently gives way to the unusual in an observer's eyes. Truth is often lost in the unrecognized pursuit of exceptions.

An unbiased answer to our question on assortative mating will be provided by studying the association in a *random* sample of the general population. Such a sample is presented in Fig. 17.[1] Stature of the wife is represented on the scale of ordinates; stature of the husband is given by the abscissal scale. Since stature is truly continuous and proceeds by intangible increments, the entire surface represents possible combinations of magnitudes of the two variables.

[1] The scatter presented in Fig. 17 is derived from data published in another form by Karl Pearson and Alice Lee in a study entitled "On the Laws of inheritance in man. I. Inheritance of physical characters." *Vide* Biometrika, 2: 357–462. 1903.

FIGURE 17

SCATTER DIAGRAM OF ASSORTATIVE MATING FOR STATURE IN 1079 HUSBANDS AND WIVES[1]



In the absence of assortative mating for stature, the dots representing actual matings would not tend to cluster about any line of trend. Neither should they be expected to be equally distributed over the entire surface, for all statures do not occur with equal frequency. Stature in each sex is distributed very nearly in accordance with the normal curve. Hence, in the bivariate surface for assortative mating the clustering would be expected, in the *absence* of likeness in stature of the matings, to occur about a central point given by the intersection of the two means, and to fall away symmetrically about this point. From a knowledge of the

univariate distributions, one may readily construct in advance the bivariate surface for complete absence of association between the two variables. It must be quite clear from Fig. 17 that, far from there being a "law of opposites" governing human matings, there is rather a tendency for marriage to take place between individuals of similar stature.

The correlation table formed by grouping the frequencies within the class boundaries indicated in Fig. 17 is presented as Fig. 18. The superiority of the scatter diagram in portraying the association will be evident after comparison of these two presentations. But whereas the scatter diagram is most useful optically, the table is necessary for computational work.

FIGURE 18

CORRELATION TABLE FOR ASSORTATIVE MATING FOR STATURE

(Data of Figure 17)

Stature of wife in inches	59.5	60.5	61.5	62.5	63.5	64.5	65.5	66.5	67.5	68.5	69.5	70.5	71.5	72.5	73.5	74.5	75.5	
	2	5	11	32	43	77	122	146	141	162	126	101	57	32	15	5	2	1079
70.5												1		1				2
69.5											1	2		2				5
68.5									1	2	1	1	1	1	1			8
67.5								2	2	3	3	2	2	1	1			16
66.5				1	1	1	3	5	4	7	8	7	3	2	1	1		44
65.5				1	2	3	5	10	8	15	13	11	6	4	2			80
64.5		1	1	3	6	7	10	16	18	24	17	13	7	4	2	1	1	131
63.5			3	3	4	9	21	15	20	19	24	23	11	4	3	2		161
62.5		1	2	4	8	10	26	22	28	28	23	14	11	4	2	1	1	185
61.5	1			8	8	15	17	25	20	24	14	14	10	4	1			161
60.5		1	2	4	5	18	16	24	15	23	13	8	5	4	1			139
59.5	1	1	1	4	4	8	10	15	16	9	5	3	1	1	1			80
58.5		1	1	1	2	4	7	6	5	4	3	2						36
57.5			1	1	2	2	4	3	3	3	1							20
56.5				1	1		1	2	1									6
55.5				1						1								2
54.5							1											1
53.5								1										1
52.5							1											1

Stature of husband in inches

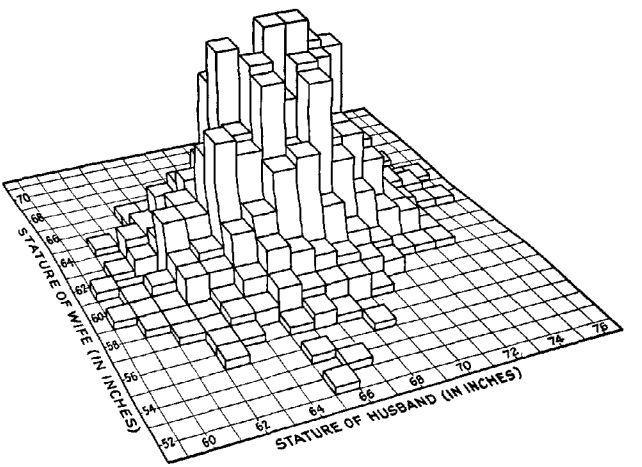
The representation of this bivariate surface may now conveniently be carried a step further to the three-dimension figure. This may be achieved by erecting above the "cells" of the table in Fig. 18 a series of rectangular prisms, each of a height proportionate to the frequency of the cell which forms its base. This is correct procedure since in the seriation all classes have the same range for each variable. Frequency now has the dimensions of volume; it is spread over an area, and the volumes will be proportional to height only if the cross section remains constant.

A diagram of a frequency solid so prepared for the assortative mating data is given as Fig. 19. The correspondence between this solid and the histograms of the univariate systems is immediately apparent. One may refer to Fig. 19 as that of a solid histogram. Just as the univariate histograms may be smoothed by curves, so the solid histograms may be graduated by smooth curving surfaces more appropriately representing the infinitely large populations from which samples are drawn. Such a graduation has been made for Fig. 19, and on the smooth curved surface

FIGURE 19

PERSPECTIVE VIEW OF THE FREQUENCY SOLID FOR ASSORTATIVE MATING IN MAN FOR STATURE

(Data of Figure 18)



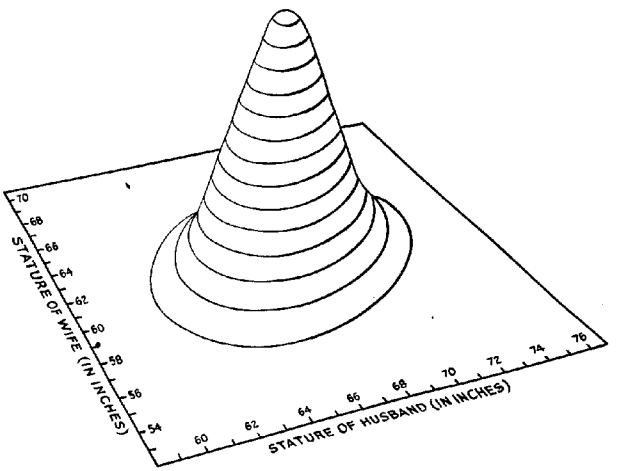
contour lines, or *isograms*, have been drawn joining the points of equal height above the base. The solid in perspective and its top elevation showing only the contours are reproduced in Figs. 20 and 21, respectively.

The remarkable and beautiful simplicity of the representation of a bivariate distribution of frequency given in Fig. 21 must surely appeal to all. It is precisely analogous to that first uncovered by Francis Galton in his study of inheritance of stature in man, which provided the foundation upon which his brilliant pioneering mind built an enduring edifice. Just

as the scatter diagram portrays any actual bivariate frequency distribution satisfactorily in two dimensions, so the contour diagram may trace the same data in graduated and more readily assimilable form. Different intensities of association between paired variables will show up with remarkable clarity in terms of such contour diagrams. Complete systems characterized by common elements spring immediately into view, and it was upon one such system that Francis Galton focused his analytical power. It was the system defined by the *normal surface* in its transition from no correlation at all to perfect association of the two variables.

FIGURE 20

PERSPECTIVE VIEW OF THE NORMAL FREQUENCY SOLID GRADUATING FIGURE 19



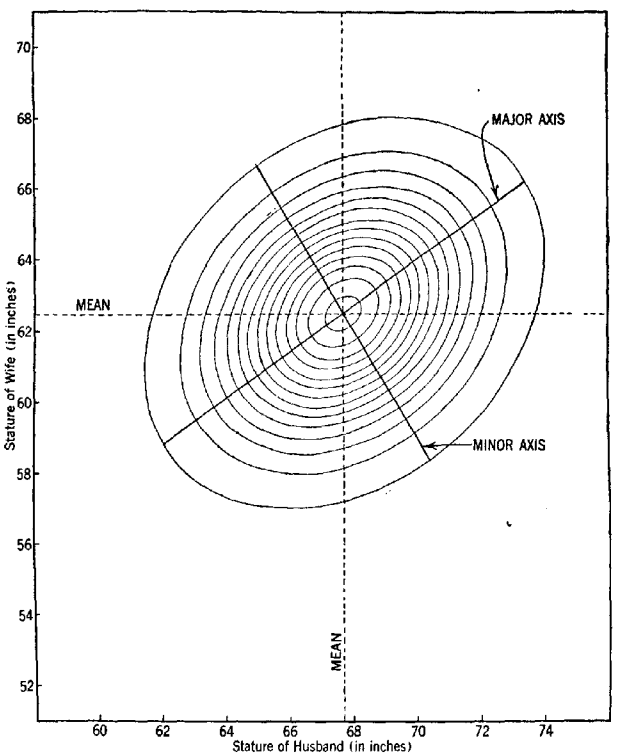
In drawing the foregoing illustrations we have assumed the assortative mating data to belong to this system, and to the accuracy of that assumption reasonable challenge is not likely to be given. The normal bivariate frequency distribution is characterized by two readily recognizable and simple features:

(1) normal distribution of both variables considered singly, and
(2) straight-line average dependence of one variable on the other.

The univariate distributions for stature in husband and wife, which may readily be plotted from the marginal columns of Fig. 18, are graduated very well by normal curves. Questions of the average interdependence of these two variables raise a new issue. On the average, how does the

FIGURE 21

CONTOUR DIAGRAM OF THE NORMAL SURFACE FITTED TO THE DATA OF FIGURE 17



stature of the wife change with that of the husband? The small circles in Fig. 17 locate for the observed data the average stature of wife within each successive inch range of stature of husband. The *trend* is obviously of a straight-line nature. A corresponding plot for change in husband's stature with that of the wife, given by the small crosses, shows the same

straight-line trend. Determination of these graduating straight lines will be reserved for the next chapter; meantime we are concerned solely with the demonstration of rectilinear average dependence of each variable on the other given by the located bands of averages.

Just as the ellipse is to be regarded as the transitional figure between a circle and a straight line, so the elliptical contours of Fig. 21 may be regarded as representing an intermediate stage between total absence of association (given by a system of circular contours) and perfect dependence between two variables given by a straight-line relationship. Surely this degree of transition may be expressed by a pure number on a simple scale! So the concept of a coefficient of correlation was nurtured in Galton's mind. A measure of partial association! The prospect this opened to his searching vision is aptly portrayed in his words:

> This part of the inquiry may be said to run along a road on a high level, that affords wide views in unexpected directions, and from which easy descents may be made to totally different goals. . . . I have a great subject to write upon, but feel keenly my literary incapacity to make it easily intelligible without sacrificing accuracy and thoroughness.

The correlation scale selected by Galton covered the limited range from zero to unity, the coefficient bearing a positive or negative sign as the one character varied directly or inversely with its correlated character. This scale of elemental simplicity has persisted without challenge. A coefficient of zero means that there is no interrelationship at all between the two variables. Knowledge of the magnitude of one does not lead to any knowledge whatsoever of the largeness or smallness of the other. On the other hand, a correlation coefficient of unity means that when the measure of one character is known the other is also fully defined. Intermediate values designate degrees of intensity of association within these natural limits.

DERIVATION OF A MEASURE OF ASSOCIATION

Let x and y designate two associated variables for which it is desired to measure the intensity of correlation. They may represent two measures of different characters on the one individual, or measures of the same or different characters on different individuals, but they should certainly arise for consideration as pairs for some logical reason. If the number of such paired observations is N, then the series may be described algebraically as

$$x_1, x_2, x_3, \cdots, x_N,$$
$$y_1, y_2, y_3, \cdots, y_N.$$

For purposes of giving explicit form to our general concept of correlation, let it be assumed by way of example that y has a perfect positive correlation with x. By this we mean that a large value of x has associated with it a *correspondingly* large value of y, and *vice versa*. The notion needing explicit definition is that of corresponding largeness or smallness. The magnitude of each individual measurement is to be evaluated for its relative largeness or smallness in comparison with the whole set of measures to which it belongs. All the necessary descriptions to enable this are at hand. The relative deviate measures precisely the relative largeness of the individual in its own group.

The scale of relative deviates has already received some consideration. We have observed its pure number character, giving universal comparability to its units. Also, the mean and standard deviation of the relative deviates comprising any group of variates are always zero and unity, respectively. Correlation analysis brings us now to paired values of relative deviates, each member of the pair coming from its own group or variable.

Transferring symbolism to the k notation used previously, wherein

$$k_x = \frac{x - \bar{x}}{s_x}, \quad \text{and} \quad k_y = \frac{y - \bar{y}}{s_y},$$

it will be recognized immediately that any system of N paired values in which k_x equals k_y *within every pair* is a system of perfect correlation. Whatever the magnitude of one variable may be, the associated value of the other has precisely the same relative magnitude. If the signs are alike within the pairs, the correlation is spoken of as positive; if they are opposed, it is designated as negative. The problem of measuring correlation in general, as one may appreciate it at this point, may be considered as a problem of measuring the "degree of likeness" in magnitude and sign of the members of the N pairs of relative deviates which may be computed from the N pairs of variates. From N pairs of relative deviates a single value is to be derived having the following properties. If the agreement between these two values is perfect in every pair, then it is desired to describe the association by a numerical value of unity. If there is no association whatever, then a value of zero on the scale must be awarded. Intermediate degrees of association must be portrayed by values between these two limits. If the paired relative deviates tend to have the same sign, the association may be designated as positive; if unlike, then negative.

It is an obvious prerequisite that, if a single numerical quantity descriptive of the correlation between a set of N paired magnitudes is to be

derived from those N pairs, then the pairs must first be fused by some appropriate method before an average is secured. The simplest method of fusion which proves fruitful is that of multiplying the paired relative deviates together. Adding these products and dividing by N to secure an average, we have

$$\frac{\Sigma k_x k_y}{N} = \frac{\Sigma\left[\frac{x-\bar{x}}{s_x}\right]\left[\frac{y-\bar{y}}{s_y}\right]}{N}$$

$$= \frac{1}{s_x s_y}\left[\frac{\Sigma(x-\bar{x})(y-\bar{y})}{N}\right]. \qquad (1)$$

The standard deviations, being constants, may be placed outside the summation sign, leaving the mean product of the deviates as the primary variable in the function. In analyzing the descriptive value of the mean product of the relative deviates, one may advantageously study first the mean product of the deviates, then later divide by the term $s_x s_y$.

The change in nature of the mean product of the deviates may be scrutinized in a very simple geometric manner. Frequency surfaces of different degrees of correlation may be represented by means of elliptical contours, one only being sufficient for each degree of correlation considered. In Fig. 22 increasing degrees of correlation are portrayed in the diagrams arranged in order from (A) to (F), the former designating zero correlation, and the latter, perfect correlation, between the two variables. Lines corresponding to the mean values, $\bar{x}$ and $\bar{y}$, divide each of these surfaces into two sections if each mean is considered alone, or into quadrants when both are considered. All values of x to the right of the vertical mid-line will have a positive sign for $x - \bar{x}$; those to the left will have a negative sign for the deviation. Similarly, all values of y above the horizontal mid-line will be positive deviates; those below, negative deviates.

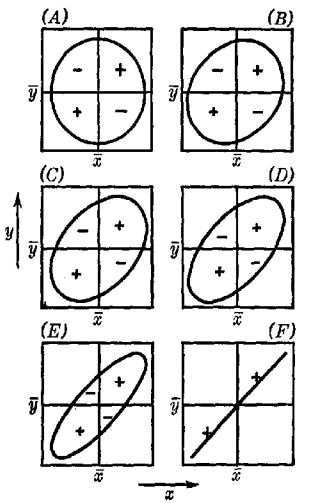
If each quadrant is now considered individually, it will be clear that the product of all the paired deviates within it must have the same sign. As one proceeds from the upper right quadrant in a clockwise direction, the sign of the product in successive quadrants will be plus, minus, plus, and minus. Thus, opposite quadrants have the same sign for their products of deviates. The *total* product of the deviates for each of the successive surfaces in Fig. 22 may be judged for its nature in a relative way by visual summation in terms of signs and areas of these four parts in each surface.

The symmetry of surface (A), portraying complete absence of correlation between the two variables, clearly makes the total product zero in

that case, and therefore equation (1) takes the value zero. However, as the degree of correlation increases, so the proportion of the correlation surface in two opposite quadrants increases at the expense of that in the other pair of quadrants, and the total product will increase in magnitude to higher and higher values. But that is not all! When the association between x and y is of a negative character, the ellipse must traverse the surface in a direction at right angles to that given in Fig. 22, and accordingly the sign of the total product will become negative while its absolute

FIGURE 22

PROGRESSIVE CHANGE IN THE TOTAL PRODUCT MOMENT, $\Sigma(x-\bar{x})(y-\bar{y})$, WITH INCREASING INTENSITY OF CORRELATION IN NORMAL SURFACES



magnitude remains unchanged. It follows directly, therefore, that the mean product of the deviates for normal surfaces is zero in the absence of correlation, and advances in magnitude away from this point progressively in a positive or negative direction as correlation of an increasingly positive or negative character is encountered.

The limiting form in this transition of ellipses is given by the straight line in surface (F), representing perfect correlation between the two variables. Now equation (1) clearly takes a minimum limiting value of zero for surface (A). Does it also assume a finite value for the other limit given by surface (F)?

For perfect correlation of the character defined, the paired relative deviates must equal one another. Therefore,

$$\frac{\Sigma\left[\frac{x-\bar{x}}{s_x}\right]\left[\frac{y-\bar{y}}{s_y}\right]}{N} = \frac{\Sigma\left[\frac{x-\bar{x}}{s_x}\right]^2}{N}$$

$$= \frac{1}{s_x^2}\frac{\Sigma(x-\bar{x})^2}{N}$$

$$= \frac{s_x^2}{s_x^2}$$

$$= 1.$$

Let us designate this mean product of relative deviates as r. Then the lower and upper limiting values of the correlation scale are given by r when normal correlation is absent or perfect, respectively. Also, the inference that r moves between these limits progressively with smooth transition in the degree of correlation from nothing at all to perfection seems undoubtedly correct. It is more satisfying, of course, to have explicit proof of the correctness of this inference. Such proof is given in the Appendix to this chapter.

When Galton first arrived at an equivalent form of the quantity which we have symbolized as r, he named it the "coefficient of reversion." Although he recognized fully that he was analyzing the system of normal surfaces of various degrees of correlation, he was thinking primarily in terms of his problem in natural inheritance; he had not at that time grasped the more general applicability of the coefficient he had devised. Its magnitude provided a measure of the extent to which offspring tended to regress toward mediocrity, and that was the immediate problem. The symbol r for this measure of reversion was selected by him and has been retained from that time. For some years it was known as "Galton's coefficient," but now the more general term "coefficient of correlation" is commonly applied to it. Galton's reasoning leading to its establishment followed an entirely different path from that which we have followed, Karl Pearson publishing in 1895 the formula

$$r_{xy} = \frac{\Sigma k_x k_y}{N}$$

$$= \frac{1}{s_x s_y}\left[\frac{\Sigma(x-\bar{x})(y-\bar{y})}{N}\right]. \qquad (2)$$

CALCULATION OF THE CORRELATION COEFFICIENT

It would be a tedious matter to compute a correlation coefficient using either of the above formulas without any modification. Deviates are numerically awkward quantities, and relative deviates would in addition be quite a nuisance to calculate. Transformation of the equations to variate form, or "moments about zero" instead of "moments about the means," is quite simple and provides a much more satisfactory formula for calculation purposes. The steps in algebraic rearrangement are as follows:

$$\frac{\Sigma(x-\bar{x})(y-\bar{y})}{N} = \frac{\Sigma(xy - x\bar{y} - \bar{x}y + \bar{x}\bar{y})}{N}$$

$$= \frac{\Sigma xy}{N} - \frac{\Sigma x\bar{y}}{N} - \frac{\Sigma \bar{x}y}{N} + \frac{\Sigma \bar{x}\bar{y}}{N}$$

$$= \frac{\Sigma xy}{N} - \bar{y}\frac{\Sigma x}{N} - \bar{x}\frac{\Sigma y}{N} + \bar{x}\bar{y}$$

$$= \frac{\Sigma xy}{N} - \bar{y}\bar{x} - \bar{x}\bar{y} + \bar{x}\bar{y}$$

$$= \frac{\Sigma xy}{N} - \bar{x}\bar{y}.$$

Therefore,

$$r_{xy} = \frac{\dfrac{\Sigma xy}{N} - \bar{x}\bar{y}}{s_x s_y}. \qquad (3)$$

This formula, suggested by Harris,[2] is admirably adapted to practical calculation needs.

(A) Calculation without Seriation

Equation (3) above permits of expeditious computation of the correlation coefficient without formation of a correlation table. The necessary summations may be proceeded with by machine directly from the original data. Illustration of this may be given in terms of the data presented as Table 13, which shows the records[3] of body weight and oxygen consump-

[2] J. Arthur Harris. The arithmetic of the product moment method of calculating the coefficient of correlation. Am. Nat., 44: 693–699. 1910.

[3] J. Arthur Harris and Francis G. Benedict. A biometric study of basal metabolism in man. Carnegie Institution of Washington. 1919.

tion under basal conditions for 136 men. The coefficient of correlation between the two variables may be determined directly from these data by means of equation (3). The necessary summations of x, x^2, y, y^2, and

TABLE 13

DATA FOR STUDY OF THE RELATIONSHIP BETWEEN WEIGHT OF MAN AND OXYGEN CONSUMPTION UNDER BASAL CONDITIONS

(Data from Harris and Benedict[3])

x = body weight in kilograms. N = 136.
y = oxygen consumption in cubic centimeters per minute.

x	y	x	y	x	y	x	y	x	y
74.0	267	88.5	289	79.0	276	78.9	302	82.1	291
82.2	282	108.9	361	62.4	259	66.0	242	74.0	270
71.2	259	73.9	264	63.5	228	63.5	238	56.7	218
56.3	223	75.0	242	64.7	217	60.0	219	59.2	231
58.2	200	50.0	165	49.3	198	55.2	193	50.6	213
55.4	224	59.3	211	85.8	265	83.1	258	83.0	237
82.1	236	78.0	262	75.9	273	75.0	263	74.5	269
74.4	244	74.2	256	73.7	218	73.1	261	73.0	222
71.1	244	69.1	234	68.4	219	67.9	280	66.4	238
66.0	241	65.1	222	65.0	227	64.0	240	63.7	237
63.2	216	62.7	227	62.6	220	62.3	213	62.0	239
60.8	209	60.6	225	60.5	250	60.5	244	60.4	229
60.4	223	60.5	214	60.1	251	60.1	233	60.0	229
59.8	245	59.7	251	58.5	193	58.0	233	57.8	192
57.8	213	57.2	232	57.1	220	57.1	216	56.8	244
56.7	238	56.3	232	56.1	219	55.1	203	54.9	232
54.3	233	53.9	207	53.6	210	52.2	219	51.4	234
50.2	203	49.3	229	48.5	185	46.3	173	92.0	273
77.4	294	73.3	247	71.3	265	70.0	269	69.7	285
69.5	248	68.2	237	67.2	247	66.7	228	66.5	261
65.8	265	64.7	247	64.4	226	64.3	240	61.4	197
59.8	208	58.2	233	57.6	208	57.4	215	56.9	215
56.4	201	55.9	209	54.5	207	54.0	206	49.7	208
49.2	187	33.2	142	83.1	284	74.0	225	71.9	225
69.7	280	68.6	258	68.2	236	67.3	240	65.3	234
63.6	289	61.6	233	61.4	225	60.5	210	58.9	208
57.9	220	55.0	211	53.4	211	74.8	227	60.6	225
								60.3	203

xy, together with the subsequent calculations, are assembled in Table 14, which incidentally reproduces an arrangement of a work sheet which may appeal to the reader as a suitable standard form.

The above procedure of direct summation from basic records without seriation obviously depends for its usefulness on the total frequency being sufficiently small to make seriation hardly worth while as a means of shortening calculation. A machine for accumulating products is, of course, essential. A scatter diagram or correlation table might well have to be prepared as an adjunct, however, if it is desired to examine the form

TABLE 14

CALCULATION SHEET FOR THE COEFFICIENT OF CORRELATION BETWEEN THE VARIABLES OF TABLE 13

x = body weight. y = oxygen consumption. N = 136.	Key: H. and B. Adult men.
$\Sigma x = 8717.1.$	$\bar{x} = 64.096.$
$\Sigma x^2 = 573{,}149.09.$	$\frac{\Sigma x^2}{N} = 4{,}214.332.$
$s_x^2 = 105.993.$	$s_x = 10.295.$
$\Sigma y = 31{,}818.$	$\bar{y} = 233.96.$
$\Sigma y^2 = 7{,}561{,}352.$	$\frac{\Sigma y^2}{N} = 55{,}598.18.$
$s_y^2 = 862.82.$	$s_y = 29.37.$
$\Sigma xy = 2{,}072{,}134.9.$	$\frac{\Sigma xy}{N} = 15{,}236.28.$
$\frac{\Sigma xy}{N} - \bar{x}\bar{y} = 240.574.$	$s_x s_y = 302.412.$
$r_{xy} = +0.796.$	

of the bivariate distribution and the nature of the average dependence lines. In some situations, of course, one knows well enough from past experience with the variables concerned that the frequency surface is reasonably normal in form. Under such conditions formation of a correlation table may be unnecessary and the above procedure may be followed.

(B) Calculation from a Correlation Table

To proceed to calculate a correlation coefficient as a measure of association without any knowledge of the type of bivariate distribution being dealt with is to court trouble. The correlation table provides a suitable basis of judgment as to the characteristics of the surface, at least as far as reasonable normality of distribution is concerned. Preparation of such a

table consumes a little time, but this will be offset in general by speedier calculation procedure if the total frequency is very much in excess of 100, and particularly if the variate magnitudes run to several digits. The data of Table 13, for instance, might well be arranged to advantage into a correlation table by an inexperienced computer. When it is recalled that seriation with uniform grouping opens the door at once to the use of code

TABLE 15

A SERIATION WITH BROAD GROUPING FOR THE DATA OF TABLE 13

x	0	1	2	3	4	5	6	7		
f	1	6	43	52	24	8	1	1	136	
360–379								1	1	11
340–359										10
320–339										9
300–319					1				1	8
280–299				4	1	4			9	7
260–279				2	11	1	1		15	6
240–259			3	13	6	1			23	5
220–239		1	13	23	4	2			43	4
200–219		1	23	9	1				34	3
180–199		3	3	1					7	2
160–179		1	1						2	1
140–159	1								1	0
	30–39.9	40–49.9	50–59.9	60–69.9	70–79.9	80–89.9	90–99.9	100–109.9	f	Y

Oxygen consumption, y

Body weight, x

scales, enabling quick computation even without a calculating machine, the advantages of the correlation table become considerably enhanced.

A correlation table may be prepared expeditiously from record cards once the grouping scheme for each variable is decided. The cards may be seriated first for one variable, then reseriated for the second variable within the classes of the first. The frequencies may be entered progressively in a blank correlation table previously prepared. The seriations

[Continue on page 102]

TABLE 16

TABULAR ORGANIZATION OF SUMMATIONS FOR SECURING THE CORRELATION COEFFICIENT

(Data of Table 15)

(12)	$x\Sigma y_x$	0	14	286	678	520	235	36	77	1,846					
(11)	Σy_x	0	14	143	226	130	47	6	11	577					
(10)	fx^2	0	6	172	468	384	200	36	49	1,315					
(9)	fx	0	6	86	156	96	40	6	7	397					
(8)	x	0	1	2	3	4	5	6	7						
(7)	f	1	6	43	52	24	8	1	1	136		577	2,765	397	1,846
Oxygen consumption, y	369.5								1	1	11	11	121	7	77
	349.5										10				
	329.5										9				
	309.5					1				1	8	8	64	4	32
	289.5				4	1	4			9	7	63	441	36	252
	269.5				2	11	1	1		15	6	90	540	61	366
	249.5			3	13	6	1			23	5	115	575	74	370
	229.5		1	13	23	4	2			43	4	172	688	122	488
	209.5		1	23	9	1				34	3	102	306	78	234
	189.5		3	3	1					7	2	14	28	12	24
	169.5		1	1						2	1	2	2	3	3
	149.5	1								1	0	0	0	0	0
		34.95	44.95	54.95	64.95	74.95	84.95	94.95	104.95	f	y	fy	fy^2	Σxy	$y\Sigma xy$
		Body weight, x								(1)	(2)	(3)	(4)	(5)	(6)

Coding constants: For x, $a = 34.95$, and $b = 10$.
For y, $a = 149.5$, and $b = 20$.

EXPLANATION OF TABLE 16

In Table 16 the extension columns (3) and (4), also (9) and (10), provide the sums within classes for the first and second moments for x and y. Their totals are transferred as entries in Table 17.

Columns (5) and (11) are formed in each case from accumulation of the products of the cell frequencies in the array and the variable scale values *parallel* to it. It gives the sum of the x values for all individuals having a fixed y value, and *vice versa*. Designations of the type Σx_y are therefore used. By way of illustration, the vertical array of total frequency 8 may be located in the correlation table. All these individuals have a body-weight value of 5 units on the code scale, or 84.95 pounds in original units of measurement. Their *total* oxygen consumption in code scale units is $(4 \times 7) + (1 \times 6) + (1 \times 5) + (2 \times 4)$, or 47 units as entered in column (11). Cell entries in column (11) give such totals. Those in column (5) give the total body weight for each class of oxygen consumption. The totals of these columns, (5) and (11), must agree with the totals of columns (9) and (3) respectively if the calculations are correct, as may readily be seen by considering just what the totals describe.

Multiplication of each entry in column (5) with the corresponding entry in column (2) yields the total product of x and y values for the array of reference in the correlation table. These products form column (6), and the analogous products working the other way across the table form column (12). The sum of column (6), giving the total "product moment" for the whole surface, should, of course, be identical with the sum of column (12). The duplicate result provides an independent check. This total is transferred to Table 17 as Σxy.

If a calculating machine is being used, columns (4), (10), (6), and (12) need not be included in the table extensions. The sum of each of these columns is all that is needed, and each may be accumulated directly on the machine from the contributory multiplications. This reduces the extension to the correlation table to three columns in each direction, the first of which is written in directly, leaving only the other two each way to be calculated.

This systematic procedure in securing the summations seems to the writer to be elegant in its simplicity and in its provision of checks. It may be mastered in a single trial, and completely independent verifications are available for Σx, Σy, and Σxy. Moreover, if the entries in column (5) are divided by those in column (1), the mean values of x for the successive classes of y are available for examination of the form of average dependence. Likewise the quotients of entries in columns (11) and (7) give the average dependence of y on x.

should always be checked at each step, then finally verified by reseriating the variables in reverse order. In the absence of cards the tally-stroke method[4] of Table 5 may be used, tallying directly into the appropriate "cells" of the correlation table.

A seriation of the data of Table 13 into very broad groups is given in Table 15. Code scales have been inserted directly alongside the marginal frequency columns. If the correlation coefficient between the two variables is determined now through use of the code scales, what transformation equation will be necessary to identify the coefficient with the original scales? The reader may note immediately that, unlike the mean and standard deviation, the correlation coefficient is a pure number on a universal scale and not a quantity on the scale of any x or y variable. The *correlation* between x and y is surely not affected by our coding scheme!

$$k_x = \frac{x - \bar{x}}{s_x} = \frac{b(\mathrm{x} - \bar{\mathrm{x}})}{bs_{\mathrm{x}}} = k_{\mathrm{x}} .$$

The relative deviate corresponding to each variate magnitude is just the same on the code scale as on the original scale. Therefore,

$$r_{xy} = \frac{\Sigma k_x k_y}{N}$$

$$= \frac{\Sigma k_{\mathrm{x}} k_{\mathrm{Y}}}{N}$$

$$= r_{\mathrm{XY}},$$

and

$$r_{xy} = \frac{\dfrac{\Sigma \mathrm{XY}}{N} - \bar{\mathrm{X}}\bar{\mathrm{Y}}}{s_{\mathrm{x}} s_{\mathrm{Y}}} . \tag{4}$$

Five total "moments about zero" need to be derived from the correlation table in application of this formula. They are: $\Sigma \mathrm{X}$, $\Sigma \mathrm{Y}$, $\Sigma \mathrm{X}^2$, $\Sigma \mathrm{Y}^2$, and the total "product moment" $\Sigma \mathrm{XY}$. Table 16 sets forth all the needed steps in securing these total moments. The work sheet giving the derived statistics on the code scales, with final transformation to the true scales, is reproduced as Table 17.

[4] *Vide* page 22.

TABLE 17

CALCULATION SHEET FOR DETERMINING THE COEFFICIENT OF CORRELATION FROM THE DATA OF TABLE 16

x = body weight (in kilograms). y = oxygen consumption (in cubic centimeters per minute). $N = 136$.		Key: H. and B. Adult men.
$\Sigma X = 397$	$\bar{X} = 2.9191$	$\bar{x} = 64.141$
$\Sigma X^2 = 1315$	$\frac{\Sigma X^2}{N} = 9.6691$	
$s_X^2 = 1.1479$	$s_X = 1.0714$	$s_x = 10.714$
$\Sigma Y = 577$	$\bar{Y} = 4.2426$	$\bar{y} = 234.353$
$\Sigma Y^2 = 2765$	$\frac{\Sigma Y^2}{N} = 20.3309$	
$s_Y^2 = 2.3308$	$s_Y = 1.5267$	$s_y = 30.534$
$\Sigma XY = 1846$	$\frac{\Sigma XY}{N} = 13.5735$	
$\frac{\Sigma XY}{N} - \bar{X}\bar{Y} = 1.1887$	$s_X s_Y = 1.6357$	$r_{xy} = +0.727$

GROUPING EFFECTS

Seriation into the very broad classes selected in Table 15 was made primarily for the purpose of economizing space. It serves incidentally to demonstrate the effects of grouping on statistics. It will be noted that the correlation coefficient derived from the correlation table is +0.727, whereas that secured from the ungrouped data is +0.796. This discrepancy is not due to coding but is a result of using very coarse grouping in the table. The values for the means, the standard deviations, and the correlation coefficient in Tables 14 and 17 are assembled in Table 17a. It will be observed that, though all show some shift in magnitude, the means are affected very little. Indeed, the shift in the means is quite inconsequential. On the other hand, the standard deviations are increased approximately 4 per cent, and the correlation coefficient is decreased nearly 10 per cent.

In general it may be shown that grouping introduces no bias or other effect of consequence on the mean, but that it does raise the standard deviation, this increase causing the decrease in the correlation coefficient. A correction factor for adjustment of the standard deviation to allow for this grouping effect has been developed. It is known as "Sheppard's

correction." Its use calls for discriminating judgment of a sort rather beyond attainment by means of elementary discussion, and the writer will therefore not attempt to present it. Situations that commend broad grouping are in general situations that will benefit rather than suffer any loss through the standard deviation being somewhat higher and the correlation coefficient somewhat lower than the values to which finer classifi-

TABLE 17a

COMPARISON OF STATISTICS DERIVED FROM GROUPED AND UNGROUPED DATA

(Statistics from Tables 14 and 17)

Statistic	Value without grouping	Value with coarse grouping	Grouping effect
$\bar{x}$	64.10	64.19	Increase 0.1 per cent
$\bar{y}$	234.0	234.9	" 0.4 " "
s_x	10.30	10.71	" 4.0 " "
s_y	29.37	30.53	" 3.9 " "
r_{xy}	+0.796	+0.727	Decrease 9.5 " "

cation would lead. It is better to establish a grouping that is reasonably fine than to apply any theoretical adjustment to the correlation coefficient. Grouping gives a slight conservative bias which is not without its advantages.

THE NORMAL BIVARIATE FREQUENCY DISTRIBUTION

The correlation coefficient assumes its maximum utility as a descriptive statistic when the bivariate frequency distribution for which it is calculated is normal in form. Indeed, it is only under such circumstances that r_{xy}, when added to knowledge of $\bar{x}$, $\bar{y}$, s_x, and s_y, completes description of the association in all its details. This may be appreciated through inspection of the equation for the normal surface:

$$w = Ce^{-\frac{1}{2}k^2}, \tag{5}$$

wherein w = the ordinate at any point (x, y),

$$C = (2\pi s_x s_y \sqrt{1 - r_{xy}^2})^{-1}, \text{ and}$$

$$k^2 = \left[\frac{x - \bar{x}}{s_x}\right]^2 + \left[\frac{y - \bar{y}}{s_y}\right]^2 - 2r_{xy}\frac{(x - \bar{x})(y - \bar{y})}{s_x s_y}.$$

One is not interested here in this rather complicated equation beyond noting that the two means, the two standard deviations, and the correla-

tion coefficient collectively define a normal surface. The correlation coefficient may be regarded as the link which unites two univariate normal distributions to give the surface of joint association of those two variables, *provided that surface also is normal.* Calculation of the correlation coefficient for any surface which is not normal must at best give only a partial description of the surface. Immensely useful as the correlation coefficient is when properly applied, it is by no means a universally comparable measure of association, not even when the average dependence is rectilinear. Biologists would do well to learn that this coefficient taken alone does not fully describe any association. Before embarking on discussion of such matters, one may well study further the trend lines, and the variation about them, which characterize the normal surface.

APPENDIX TO CHAPTER 7

Proof that r_{xy} May Take Values Only Between the Limits of Plus One and Minus One

Bivariate frequency systems in which the paired values of x and y have precisely the same relative deviate values are systems of perfect straight-line correlation between the two variables. All other systems must fall into either of the following categories:

(1) All (x,y) points fall on a line which is curved. This is perfect correlation of a curvilinear character.

(2) The (x,y) points are scattered about some trend line, either straight or curved. This represents imperfect correlation of some form.

In either of these situations the relative deviates of the paired x and y values cannot be identical. The greater the departure from perfect rectilinear correlation, the more in general will the paired relative deviates tend to differ from one another.

Let d designate the difference between each pair of relative deviates in a fixed sense. Then d may be defined algebraically as

$$d = \frac{x - \bar{x}}{s_x} - \frac{y - \bar{y}}{s_y}$$

$$= k_x - k_y.$$

The mean value of d, however, will *always* be zero, for the mean value of each complete set of relative deviates is zero.

$$\bar{d} = \frac{\Sigma(k_x - k_y)}{N} = \frac{\Sigma k_x}{N} - \frac{\Sigma k_y}{N} = 0.$$

From this it follows directly that, the greater the individual values of d tend to become, some being negative and some positive, the greater the variation in d. In measuring this variation by the standard deviation of d, it follows directly that

$$s_d^2 = \frac{\Sigma d^2}{N} - \bar{d}^2 = \frac{\Sigma(k_x - k_y)^2}{N}$$

$$= \frac{\Sigma k_x^2}{N} - 2\frac{\Sigma k_x k_y}{N} + \frac{\Sigma k_y^2}{N}.$$

But the mean square of any complete set of relative deviates is itself the squared standard deviation of the set, which is always unity. Also, the mean product of any complete set of paired relative deviates is r. Therefore,

$$s_d^2 = 1 - 2r_{xy} + 1$$
$$= 2(1 - r_{xy}).$$

Rearranging this equation, it follows that

$$r_{xy} = 1 - \tfrac{1}{2}s_d^2.$$

Since s_d^2 is zero only when x and y are perfectly correlated in straight-line form, and under all other conditions s_d^2 is a real positive quantity because d then becomes a variable quantity, r_{xy} must always be less than unity under both conditions (1) and (2) above.

The sign of a correlation coefficient is quite independent of the numerical value of the coefficient. It may be reversed at will by reversing the signs of the relative deviates of one variable throughout. Therefore it follows also that r_{xy} cannot be less than minus one.

The numerical scale of r for curvilinear systems may readily be inferred from the above to be from zero to some value less than one. That latter limiting value may be demonstrated by geometric drawings to fall increasingly below unity as curvilinearity increases. Indeed, the upper limit may itself be zero under certain conditions, regardless of the intensity of adherence of the scatter to the line. Thus r_{xy} has but little meaning as applied to a curvilinear system unless the upper limit of r for the system is established. For rectilinear systems, the upper limit is, of course, known to be unity.

CHAPTER 8

RECTILINEAR REGRESSION

It having been recognized that certain variables tend to move together in their magnitudes, or are correlated in their variations, there is no more natural reaction than to seek understanding of what may be expected for the magnitude of one of them when any magnitude of the other is specified. Such understanding is comparatively easy of achievement when the variables are very highly correlated and are either susceptible of experimental control or assume values which, in the natural course of events, cover a tangible range of measurement. In such cases a simple plot of associated values for the two variables immediately establishes a trend of linear form. The elementary gas laws of chemistry and the more obvious principles of mechanics owe their early formulation to just such procedures of analysis on the part of scientists who were obliged to work with relatively crude measuring devices. Deviations from perfectly linear interrelationships were ascribed to errors of measurement, and the selected graduating lines were accepted as reflections of the true relationships. It has since been learned with more sensitive measuring devices that such deviations are not always entirely errors of measurement, that the correlations are not in fact as perfect as the first statements of the laws supposed, and that residual variations may often be accounted for through consideration of other variables whose influence was not at first recognized. The ideas of A "causing" B have been replaced by others involving the concept of B being more or less highly correlated with A, but this has not detracted from the importance of the central trend line in each case. As far as A may be predicted from B *alone*, the trend line continues to provide the best estimate.

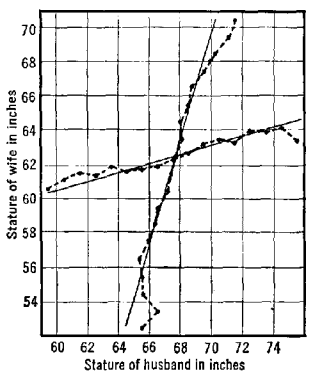
Associated variations are in general much less perfect for biological variables than those commonly found in the fields of physics, chemistry, *et cetera*, with their highly controllable experimental situations. This is naturally so, for usually a great number of variables interact to produce biological phenomena, the influence of only the more important and better understood of them being recognized. However, the consequence of very imperfect correlations does not forbid progress along the lines of determining what is the most likely value of the variable B so far as it may be determined from any magnitude of the variable A with which it

is correlated. It is to these graduating lines for predicting one variable from another that we shall now give attention. As before, analysis will be confined in this elementary study to the simplest situations, namely, those wherein the line itself is of the simplest type—a *straight* line.

To go back to Fig. 17,[1] it will be noted that, although the correlation between stature of husband and wife is far from perfect, on the average there appears to be a straight-line increase in wife's stature as husband's stature increases. The calculated averages of y in successive classes of x were plotted and joined in that diagram to reveal this feature. In analogous manner the averages of x for classes of y were determined and plotted. Both these lines, together with their rectilinear graduations,

FIGURE 23

REGRESSION LINES FOR THE DATA ON ASSORTATIVE MATING IN FIGURE 17



are reproduced in Fig. 23 apart from the scatter of individual cases. Such lines are generally known among statisticians as *regression* lines, following the terminology used by Galton in his inheritance studies. Since division of a variable into classes for purposes of getting a succession of averages to establish such a trend is a purely empirical procedure, the irregular line formed by joining the sequence of average points is referred to as an *empirical regression* line. Note, however, that though changes in the classification will change the irregularities, they will not alter the underlying trend in the data. That trend in the present illustration is most reasonably a straight line in each case. These graduating lines may be designated for purposes of distinction as the *theoretical regression* lines.

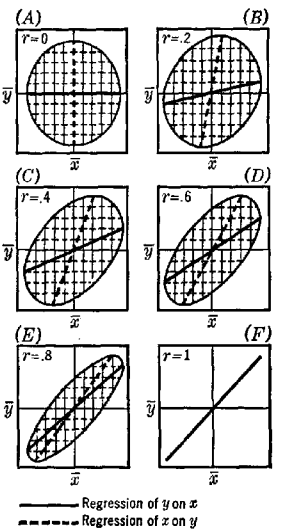
[1] *Vide* page 86.

REGRESSIONS FOR THE NORMAL SURFACE

In Fig. 24, normal surfaces of several successive degrees of correlation are represented by elliptical contours. In order that the surfaces may be directly comparable one to the other, they have been plotted in terms of the relative deviate as the unit of measurement throughout. That is, s_x equals s_y as a linear dimension in every diagram, or the "axes of symmetry" of the ellipses are at 45 degrees to the axes of abscissas and ordinates. The correlation coefficient increases from zero in panel (*A*) to unity in panel (*F*). The rectilinear regressions have been drawn in each case, and it will be observed that they accord throughout with the mid-points of arrays established in the appropriate direction.

FIGURE 24

THE CHANGE IN SLOPE OF THE REGRESSION LINES FOR NORMAL SURFACES OF PROGRESSIVELY GREATER CORRELATION



From this geometric presentation it will be seen:

(1) that straight regression lines apparently provide the correct type of graduation of array averages for normal surfaces;

(2) that there must be two regression lines for every surface of imperfect correlation;

(3) that these regression lines appear to intersect at the point $(\bar{x}, \bar{y})$ for every surface; and

(4) that the slopes of the two lines on each surface, taken with respect to their appropriate axes, are identical.

Closer study will suggest as a fifth feature that a simple relationship exists between the slope of each regression line and the magnitude of the correlation coefficient. From an inspection of panel (*A*) it will be clear that the average value of y for each value of x is the mean of y as a whole. It is a fixed value and does not change with x. Also, the mean value of x for each value of y is $\bar{x}$. The regression lines coincide with the axes of means, and are without slope in each case. As the degree of correlation increases, so these regression lines revolve away from the axes of means and toward each other until, for perfect correlation, they coincide. The slope of each of these two coincident regression lines is now unity. The slope of each regression line varies from zero to unity under these conditions as r varies from zero to unity. Does the slope coincide with the correlation coefficient in every case? We must proceed to define these lines explicitly.

Empirical regression lines for any given surface, normal or otherwise, may be established very readily through simple arithmetical procedures following classification. The method of securing the correlation coefficient detailed in Table 16[2] places one in a position to calculate the means of arrays directly, as is indicated in the explanatory legend facing the table.[3] These means may then be plotted and joined to form the empirical regression line as for the assortative mating data in Fig. 23. One may then judge whether or not a straight line will graduate the empirical averages satisfactorily. Let us assume that that judgment has been made for a particular case and the straight theoretical regression line concept is accepted for the data in hand. It is then desired to fit a *straight* line to the data. It will now be established algebraically that the line of "best fit" will pass through the intersection of means with slope r when the relative deviates form the units of measurement. In so doing, acceptable principles for determining what comprises a "best-fitting" line may be established.

THE DEPENDENT VARIABLE

The definitive terms "dependent" and "independent," as applied to variables, are most useful in discussing associations. Prior to undertaking our special problem a word or two about these terms may be help-

[2] *Vide* page 100. [3] *Vide* page 101.

ful. As applied in the present connection, they are of mathematical origin, and have their customary meaning in the sense that, from a point of view of mathematical analysis of the association, one of the variables is being regarded as dependent for its magnitude on the other. The other variable is designated as independent by contrast. This purely mathematical usage of the terms should not be construed as meaning dependence in the practical sense. Stature of wife or husband may hardly be claimed to depend on that of the other member to the alliance. In the statistical analysis of the data in Fig. 17,[4] however, we may choose to regard one variable as a function of, or dependent upon, the other in one approach, and then equally well reverse the arrangement in a second approach to establish both regression lines.

In some associations such as this, it is biologically sensible to consider either variable as a function of the other. More generally, however, one variable naturally forms the dependent one as far as logical analysis is concerned. The weight of the new-born infant may logically be regarded as a function of its mother's weight, but the inverse arrangement is devoid of any such logical meaning. Regression analysis of biological associations usually proceeds for such reasons in one direction only. The mistake should not be made, however, of losing sight of the fact that any bivariate frequency surface not characterized by perfect correlation has two regression lines, each one serving for prediction *in only one direction.*

REGRESSION IN TERMS OF RELATIVE DEVIATES

The equation to any straight line may be written in the general form

$$Y = a + bx, \tag{1}$$

where Y is the dependent variable and a and b are constants fixing the position of the line. The capital letter is used for the dependent variable here to characterize the theoretical value given by the equation, as opposed to the observed values y found associated with any particular value of the independent variable x. When x equals zero, then y equals the constant a, and therefore a fixes a point through which the line passes. Each unit of increase in x also causes b units of increase in y. Therefore the constant b fixes the slope of the line.

Let the contour in Fig. 25 designate a correlation surface to which it is desired to fit the regression of y on x as a straight line such as RL. For purposes of simplicity in derivation of the constants for this line we shall find it convenient to transform to the relative deviate scales for the

[4] *Vide* page 86.

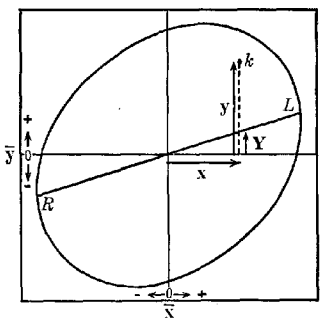
two variables. The transformation is quite comparable to that used in our coding scheme. Bold-face type will now be used for all quantities referred to these scales. The line RL may then be defined by the equation

$$\mathbf{Y} = \mathbf{a} + \mathbf{bx}, \tag{2}$$

the constants **a** and **b** being so chosen that the line is the "best-fitting" straight line that may be inscribed on the surface for the regression of y on x. Let the point k represent any individual dot in the swarm which is represented as a whole by the elliptical contour. The coordinates **x** and **y** will be appropriate to this general point k.

FIGURE 25

COORDINATES OF ANY POINT k IN A NORMAL SURFACE, SCALED IN RELATIVE DEVIATES, IN RELATION TO THE REGRESSION OF y ON x



The simplest requirement to be demanded of such a line of "best fit" would be that the positive deviations of observed values from the line should, in their totality, be equal to the negative deviations. In other words, the sum of all the deviations from the line shall be zero. If the line is to be a true moving average, then this condition must be fulfilled.

Now
$$\Sigma(\mathbf{y} - \mathbf{Y}) = \Sigma[\mathbf{y} - (\mathbf{a} + \mathbf{bx})]$$
$$= \Sigma\mathbf{y} - N\mathbf{a} - \mathbf{b}\Sigma\mathbf{x}.$$

But $\Sigma\mathbf{y}$ and $\Sigma\mathbf{x}$ are both equal to zero. Since N is a real number, then **a** must be made zero if the whole expression is to equal zero, thus fulfilling the requirement of good fit. Returning to equation (2), it will be observed that **a** is the value of **Y** when **x** is equal to zero, so that the requirement merely demands that the line shall pass through the point ($\mathbf{x} = 0$, $\mathbf{Y} = 0$), which is the same point as ($\bar{x}$, $\bar{y}$) when referred to the original scales of measurement.

Any line passing through the intersection of means fulfills the first requirement of good fit. Of this infinity of lines there is, of course, just one of "best fit," and its slope $\mathbf{b}$ must now be fixed by some further specification. It is necessary to define just what is meant by a "best-fitting" line. This line must surely be definable in some way by the statement that it is the one of closest approximation to the swarm; that is, it is the line from which the points in the swarm deviate the least. All the deviations $\mathbf{y} - \mathbf{Y}$ must be considered positive values in this connection, which may be achieved either by disregarding their signs or by using the second powers instead of the first. The much greater mathematical simplicity of solution of such problems in general when squaring is involved rather than ignoring the signs has been commented on previously in these pages. Let us then express our requirement in this technical form: The line of "best fit" is that line from which the mean square deviation of the observed values is a minimum.

Since $\mathbf{Y} = \mathbf{bx}$, $\mathbf{a}$ having been set equal to zero,

$$\frac{\Sigma(\mathbf{y} - \mathbf{Y})^2}{N} = \frac{\Sigma(\mathbf{y} - \mathbf{bx})^2}{N}$$
$$= \frac{\Sigma \mathbf{y}^2}{N} - 2\mathbf{b}\frac{\Sigma \mathbf{xy}}{N} + \mathbf{b}^2\frac{\Sigma \mathbf{x}^2}{N}.$$

But the mean square of the relative deviates, $\mathbf{x}$ and $\mathbf{y}$, is unity in each case. Also, the mean product of the relative deviates is r_{xy}. Therefore

$$\frac{\Sigma(\mathbf{y} - \mathbf{Y})^2}{N} = 1 - 2\mathbf{b}r_{xy} + \mathbf{b}^2.$$

Adding and subtracting r_{xy}^2 to form a binomial, the quantity to become minimized through selection of the appropriate value of $\mathbf{b}$ becomes

$$\frac{\Sigma(\mathbf{y} - \mathbf{Y})^2}{N} = 1 - r_{xy}^2 + r_{xy}^2 - 2\mathbf{b}r_{xy} + \mathbf{b}^2$$
$$= (1 - r_{xy}^2) + (r_{xy} - \mathbf{b})^2.$$

Clearly the objective will be achieved if $\mathbf{b}$ is set equal to r_{xy}. The regression line of "best fit by least squares" will pass through the intersection of means with slope equal to the correlation coefficient, when the variables are scaled in terms of relative deviates.

$$\mathbf{Y} = r_{xy}\mathbf{x}. \tag{3}$$

Also, the mean square deviation of the observed values about the regression line under the same conditions is

$$\frac{\Sigma(\mathbf{y} - \mathbf{Y})^2}{N} = 1 - r_{xy}^2. \tag{4}$$

THE TRUE SCALE EQUATIONS

This regression equation may be transferred from the scale of relative deviates to the original scales of measurement of the variables by direct substitution.

$$\mathbf{Y} = r_{xy}\,\mathbf{x},$$

that is $$\frac{Y - \bar{y}}{s_y} = r_{xy}\frac{x - \bar{x}}{s_x},$$

or $$Y - \bar{y} = r_{xy}\frac{s_y}{s_x}(x - \bar{x}),$$

and $$Y = \left[\bar{y} - r_{xy}\frac{s_y}{s_x}\bar{x}\right] + \left(r_{xy}\frac{s_y}{s_x}\right)x.$$

Thus $$Y = a + bx,$$

where $$b = r_{xy}\frac{s_y}{s_x}, \qquad (5)$$

and $$a = \bar{y} - b\bar{x}.$$

Similarly the rectilinear regression of x on y will be given by the equations

$$X = a + by,$$

where $$b = r_{xy}\frac{s_x}{s_y}, \qquad (6)$$

and $$a = \bar{x} - b\bar{y}.$$

It is important to recognize in both sets of equations (5) and (6) that r_{xy} is a quantity *with sign*; b will always be negative when r is negative, and a will then become the sum of two numbers.

THE REGRESSION COEFFICIENT

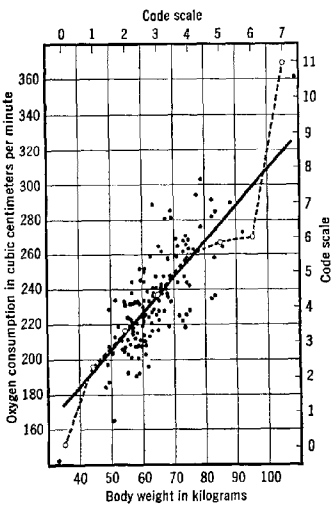
The slope b of each regression line, expressing in the concrete units of actual measurement the average change in the dependent variable for each unit of change in the other, is known as the *regression coefficient.* It is given in each case by the product of the correlation coefficient and the appropriate quotient of the standard deviations of the two variables, that of the dependent variable forming the numerator.

Even more than the correlation coefficient, the regression coefficient is a statistic of great descriptive value. Whereas the former measures intensity of association on a relative scale, the latter measures *average*

change in terms of the scales employed for the measurement of the variables. The regression coefficient is dependent for its validity solely on the existence of a straight-line average relationship of the dependent to the independent variable.

FIGURE 26

THE REGRESSION OF BASAL METABOLISM ON BODY WEIGHT FOR 136 NORMAL MEN



CALCULATION PROCEDURE

The computational steps in determining the constants for a rectilinear regression line are very simple when the correlation coefficient for the surface has previously been determined. For purposes of illustration we reproduce in Fig. 26 the scatter diagram for the association of basal metabolism with weight of subject for 136 normal men (see Tables 13 and 15 of previous chapter). One may well be concerned in ascertaining from such data the most likely oxygen consumption for the successive weights. Calculations leading to the desired empirical and theoretical regression lines are given in Table 18 with its appended work sheet. The first three columns of the table, and the means, standard deviations, and correlation coefficient, are copied from Tables 16 and 17

* For basic data, see Table 13, page 97.

respectively. The very few steps to be taken in securing the regression constants a and b may be noted in particular. The derived regression lines are given in Fig. 26, wherein the set of empirical averages is plotted by small circles which are joined by broken lines to indicate the continuity. In order to draw the theoretical regression line it is necessary to

TABLE 18

CALCULATIONS FOR THE REGRESSION OF OXYGEN CONSUMPTION ON BODY WEIGHT

(Data from Tables 16 and 17)

Empirical regression

x = body weight.
y = oxygen consumption.

Code scale values				Original scale values	
x	f	Σy_x	$\bar{y}_x$	x	$\bar{y}_x$
0	1	0	0	35	150
1	6	14	2.33	45	197
2	43	143	3.33	55	217
3	52	226	4.35	65	237
4	24	130	5.42	75	258
5	8	47	5.88	85	268
6	1	6	6.00	95	270
7	1	11	11.00	105	370

Theoretical regression

$\bar{x} = 64.141.$ $\bar{y} = 234.353.$

$s_x = 10.714.$ $s_y = 30.534.$

$r_{xy} = +0.727.$

$\frac{s_y}{s_x} = 2.850.$ $b = r_{xy}\frac{s_y}{s_x} = 2.072.$

$a = \bar{y} - b\bar{x} = 101.45.$

$Y = 101.45 + 2.072x.$

When $x = 35$, then $Y = 174.4.$
When $x = 105$, then $Y = 319.4.$

plot two points determined by its equation. Greater accuracy in plotting this line will be secured by taking two points far apart. It is therefore most useful to substitute in the equation a small and a large value of the independent variable, as has been done in Table 18.

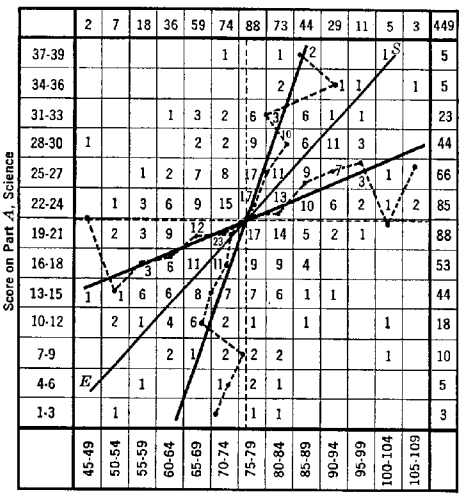
As a partial check on the accuracy of the computations, note that the line as drawn passes through the point $(\bar{x}, \bar{y})$, as is demanded by theory.

The agreement by visual inspection of the empirical and theoretical lines gives an approximate verification of the correctness of the slope of the theoretical line. Complete verification must, of course, be given by thorough checking of the arithmetic itself.

Further illustration of the application of rectilinear regression lines may be given with data from the field of educational statistics. For this purpose we shall now study the correlation between the marks secured by 449 students in the "science" and "current affairs" tests already

FIGURE 27

CORRELATION TABLE, REGRESSION LINES, AND EQUIVALENT SCORE LINE (*ES*), FOR MARKS SECURED ON TWO PARTS OF A REFLECTIVE THINKING EXAMINATION



Score on Part A, Science	45-49	50-54	55-59	60-64	65-69	70-74	75-79	80-84	85-89	90-94	95-99	100-104	105-109	
37-39						1		1	2			1		5
34-36								2	1	1			1	5
31-33				1	3	2	6	5	6	1	1			23
28-30	1				2	2	9	10	6	11	3			44
25-27			1	2	7	8	17	11	9	7	3	1		66
22-24		1	3	6	9	15	17	13	10	6	2	1	2	85
19-21		2	3	9	12	23	17	14	5	2	1			88
16-18			3	6	11	11	9	9	4					53
13-15	1	1	6	6	8	7	7	6	1	1				44
10-12		2	1	4	6	2	1		1			1		18
7-9				2	1	2	2	2				1		10
4-6			1			1	2	1						5
1-3		1					1	1						3
	45-49	50-54	55-59	60-64	65-69	70-74	75-79	80-84	85-89	90-94	95-99	100-104	105-109	449

referred to at the close of Chapter 6. In the earlier analysis these two distributions were related to one another for the purpose of defining scores of equivalent probability of occurrence. No consideration was given therein to the average mark *actually secured* in one test by the students receiving any given score in the other. One may now give attention to this problem.

Figure 27 presents the bivariate frequency distribution for the scores of the 449 students in the two tests under consideration. The central line *ES*, which is drawn across the frequency table as a background, is that of equivalent scores from the point of view previously set forth.

It reproduces the line of panel (*C*) in Fig. 16. The line *ES* may now be observed to form the principal axis of the swarm forming the correlation table. Defining the paired scores of equal relative deviates, it is the line on which all points would fall if there happened to be perfect rectilinear correlation between the scores secured on each test by the 449 examinees.

A glance at Fig. 27 is sufficient to show that such perfection in correlation is far from realized. What then is the most likely score on Part *B* for any given score on Part *A*, and *vice versa*? The averages of arrays suggest rectilinear graduations as satisfactory, and these are drawn as the solid lines in the figure. The broken lines parallel to the axes indicate the mean values of each examination as a whole. The imperfect correlation between the two scores has resulted in the lines of average dependence revolving away from the line of equivalent scores, *ES* (or principal axis of symmetry of the surface), toward the axes of means. That is, the most likely score on one test for a given score on the other is found over the whole surface to regress or move back from the equivalent score toward the general average of all scores, or toward mediocrity. This phenomenon is, of course, a consequence of imperfect correlation, the relative degree of regression being inversely proportional to the correlation coefficient.

It was in studying the inheritance of stature in man that Galton first found that adult sons on the average attained statures intermediate between the father's stature and the general population average. His correlation coefficient aimed to measure this degree of "reverting back" and he called it the "coefficient of reversion" as we have already noted. He also called the line tracing this average relationship on the scales of measurement the "regression line," a term which has persisted quite appropriately, for there is always relative regression on the average toward mediocrity with imperfect correlation between variables. Translating the phenomenon to the inheritance of intelligence, it becomes at once a source of hope and despair respectively for the moron and the genius; their offspring will tend on the average to be intermediate between them and the general population level in intelligence, for the parent-offspring correlation in intelligence falls far short of perfection.

CHAPTER 9

RESIDUAL VARIATION

While the regression equation permits of the determination of the most likely value of a dependent variable for any specified value of the other, it will be recognized that this is average expectancy only, based on a known scatter of individual cases. Just as the fixed value $\bar{y}$ is the mean of a univariate frequency distribution, so the line defined by the equation for Y traces the mean of a frequency distribution which is always parallel to the y axis but which moves across the (x,y) plane. For any particular value of an independent variable x, Y is the most likely value of the dependent variable in the sense that it is the expected average for any large experience. Many values of y will be observed to be associated with this particular x, these y values differing among themselves. In a scatter diagram these observed values will be represented by dots scattered along the y axis corresponding to x and about the point Y on the regression line. The deviation of each of these observed values from average expectancy may be measured; it is symbolically defined in each case as $y - Y$. There will indeed be N values of $y - Y$ for any given surface, and, since a primary requirement in fitting the line was that the total deviation, $\Sigma(y - Y)$, should be zero, the mean value of $y - Y$ is, of course, zero.

The amount and kind of variation of the observed values about the regression line are of great interest. Obviously, with high correlation the dispersion from the regression line is small, vanishing entirely when correlation is perfect. As the intensity of association *decreases* from this upper limit, the points scatter *increasingly* from the axis of equivalent values. The scatter finally reaches a maximum when correlation vanishes. It is this scattering which determines the extent to which individual values of the dependent variable may be expected to deviate from the most likely value given by the regression equation.

CONSTANCY OF RESIDUAL VARIATION

For the normal surface, and indeed not infrequently for other types of bivariate distribution, the scatter about each regression line is constant in amount and form from one end of the line to the other. Some

appreciation of the feature of constancy of form may be secured from inspection of Fig. 17, wherein it may be noted that the frequency distributions of y within the successive classes of x are fairly symmetrically disposed about the regression line as a central value in each case. These distributions are indeed normal in form.

TABLE 19

CALCULATION OF "ROOT MEAN SQUARE DEVIATIONS" OF $y-Y$ WITHIN CLASSES OF x FOR THE DATA OF FIG. 27

Class x	Regression Y	Total squared deviation $\Sigma(y-Y)^2$	Frequency f	R. M. S. D. about regression line
45–49	15.3272	201.5526	2	10.0387
50–54	16.4677	330.2957	7	6.8691
55–59	17.6081	450.2102	18	5.0012
60–64	18.7486	1,010.2245	36	5.2973
65–69	19.8890	1,936.6981	59	5.7293
70–74	21.0295	2,221.5082	74	5.4791
75–79	22.1699	3,558.3835	88	6.3589
80–84	23.3104	3,345.8714	73	6.7701
85–89	24.4508	1,521.3338	44	5.8801
90–94	25.5913	520.8516	29	4.2380
95–99	26.7317	193.0290	11	4.1890
100–104	27.8722	777.2805	5	12.4682
105–109	29.0127	102.8645	3	5.8556
All classes		16,170.1037	449	6.0011

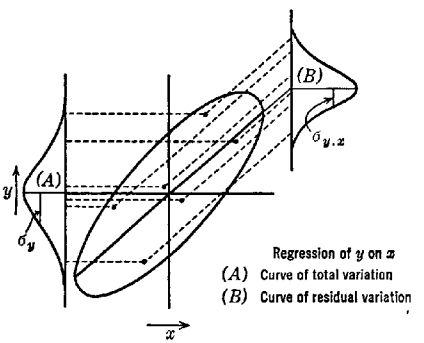
Calculations given in Table 19 of the "root mean square deviation" [1] from the regression line for the data of Fig. 27 show a fairly constant value for the amount of variation, the fluctuations being justly ascribable to the sampling errors attending the small numbers of cases. This feature of constancy of scatter about the regression line is commonly designated by the technical term *homoscedasticity* (Greek: *homos*—the same; *skedasis*—a scattering). Homoscedastic normal distribution about the regression lines characterizes normal surfaces in general, Figs. 17 and 27 being fair examples.

[1] The standard deviation is a particular "root mean square deviation" in which the deviations are taken from the mean. In our consideration herein, the deviations are considered from the regression-line values, Y, which graduate the observed array means by a straight line.

We have seen that the probability of occurrence of any specified value in a normal distribution may be determined precisely if the mean and standard deviation of the distribution are known. The regression equation provides the anticipated mean value of the dependent variable for any specified value of the independent variable. May the needed standard deviation be calculated with like simplicity? Summing the squared deviations for all classes of x as in the last line of Table 19, and dividing by $N=449$, leads to the value 6.0011 for the desired standard deviation. The reader will observe, however, that these computations are quite tedious to perform. This is not a simple method of valuing the general dispersion from the regression line.

FIGURE 28

FREQUENCY CURVES OF TOTAL VARIATION, AND RESIDUAL VARIATION ABOUT THE REGRESSION LINE, FOR A NORMAL SURFACE



THE RELATIONSHIP OF s_{y-Y} TO r_{xy}

One may profitably turn at this point to algebraic analysis. For the N values of x there are N associated y values and N pairs of derived magnitudes, Y and $y - Y$. So far as y may be predicted from x, Y is that value in each case. To this must be added an increment $y - Y$ which is not predictable from the value of x alone.[2] $y - Y$ is therefore independent of x; it represents a residual part in the total magnitude of each y value which can be accounted for only through consideration of variables other than x. The fact that y is not perfectly correlated with x indicates at once that factors other than those reflected in x are at play in determining y. The scatter about the regression line portrays these

[2] Note that this increment may be positive or negative.

residuals, $y - Y$, which may be accumulated into a single distribution by projection of the points of the swarm parallel to the regression line and onto a y axis. An attempt is made to represent this distribution in Fig. 28, where a normal surface with correlation coefficient of + 0.8 is represented by a single contour. The normal curve on the left represents the y distribution, and that on the right is the $y - Y$ distribution. Six points from the swarm represented by the contour are projected to indicate the source of each univariate frequency distribution.

The mean value of the residuals $y - Y$ is always zero. However, the variation about that mean value increases from nothing at all to a maximum magnitude as the correlation coefficient passes from unity down to zero. This variation may be measured as the standard deviation of the residuals curve in Fig. 28. Now the squared standard deviation of any distribution with zero mean is the mean square of the individuals. That is,

$$s^2_{y-Y} = \frac{\Sigma(y - Y)^2}{N}. \tag{1}$$

But $$Y = \bar{y} + b(x - \bar{x}),$$

where $$b = r_{xy}\frac{s_y}{s_x}.$$

Therefore
$$s^2_{y-Y} = \frac{\Sigma[y - [\bar{y} + b(x - \bar{x})]]^2}{N}$$

$$= \frac{\Sigma[(y - \bar{y}) - b(x - \bar{x})]^2}{N}$$

$$= \frac{\Sigma(y - \bar{y})^2}{N} - 2b\frac{\Sigma(x - \bar{x})(y - \bar{y})}{N} + b^2\frac{\Sigma(x - \bar{x})^2}{N}$$

$$= s^2_y - 2\,b\,r_{xy}s_xs_y + b^2s^2_x$$

$$= s^2_y - 2r_{xy}\frac{s_y}{s_x}r_{xy}s_xs_y + r^2_{xy}\frac{s^2_y}{s^2_x}s^2_x$$

$$= s^2_y(1 - r^2_{xy}).$$

That is, $$s_{y-Y} = s_y\sqrt{1 - r^2_{xy}}. \tag{2}$$

The tedious calculations of Table 19 are indeed unnecessary. The same result is given by a very simple equation. In Table 19 we found by long arithmetic calculation that $s_{y-Y} = 6.0011$. From equation (2),

$$s_{y-Y} = 6.4820\sqrt{1 - 0.9258^2}$$

$$= 6.0011.$$

ERRORS OF ESTIMATE

This standard deviation of what has so far been designated as the residual variation is often referred to as the "standard error of estimate." For normal homoscedastic systems it defines the normal curve of error about the regression-line value which may be expected in predicting one variate magnitude from knowledge of the other. The "probable error of estimate" will, of course, be given by 0.6745 times s_{y-Y}.

It is also common practice to use the subscript symbolism $y \cdot x$ in place of $y - Y$. This is a very useful convention, for

$$s_{y-Y} = s_{y \cdot x}$$

measures the variation in y which is independent of x; $s_{y \cdot x}$ is the standard deviation of y for any fixed value of x. Thus a form of symbolism is introduced which defines all elements concerned and is readily extensible to other statistics as well as to any number of variables in multiple association.

Just as Y defines the most likely value of y for any particular value of x, so $(Y \pm 0.6745\ s_{y \cdot x})$ defines the range within which the central 50 per cent of actually associated values of y may be expected to occur, provided the distribution about the regression line is characterized by normality and homoscedasticity. Similarly, only 5 per cent of the actual values will fall outside the range defined by $(Y \pm 1.96\ s_{y \cdot x})$.

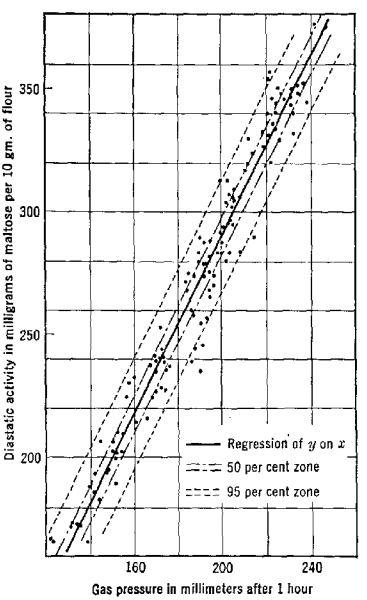
It should be noted, however, that *heteroscedasticity* is not uncommon in systems of rectilinear regression. Under such circumstances, $s_{y \cdot x}$ is only an average value for the changing scedasticity as one passes along the regression line. It will not define the standard error of estimate about any one point Y except by coincidence. After a little experience, visual inspection of the scatter diagram or correlation table will usually suffice in judging the reasonableness of the assumption of "constant scatter" for most practical purposes.

The *form* of variation about the regression line follows the normal curve quite commonly, even when the surface itself is not otherwise of the normal form. It is only in such cases, of course, that the standard error of estimate may be multiplied by factors appropriate to the normal curve for the establishment of probability zones about the regression line. In Fig. 29 an illustration is presented of the establishment of such zones. Although the regression in this surface is rectilinear and homoscedasticity prevails, the surface itself is anormal, for the univariate distributions are distinctly anormal. Nevertheless, although the correlation coefficient has doubtful descriptive value taken alone, it leads

directly to the correct regression line and error of estimate. The coarsely broken lines on either side of the regression line define the zone of "equally probable error" in the prediction of diastatic activity from carbon dioxide production. The outer finely broken lines bound the zone within which 95 per cent of all cases may be expected to occur.

FIGURE 29

RESIDUAL VARIATION ZONES ABOUT THE REGRESSION LINE IN A HOMOSCEDASTIC SYSTEM *



RELATIVE RESIDUAL VARIATION

It will be observed that the ratio of the error of estimate to the total variation in the dependent variable may be expressed as a function solely of the correlation coefficient.

$$\frac{s_{y \cdot x}}{s_y} = \sqrt{1 - r_{xy}^2}.$$

* Data from C. E. Davis and D. F. Worley. Cereal Chem., 11: 536-545. 1934.

The value of this ratio is often now called the *coefficient of alienation*, for it measures the proportion of the variation in the dependent variable which is alienated from or independent of the associated variable. This value is the same for any normal surface regardless of which variable is considered as dependent.

$$\text{C.A.} = \frac{s_{y \cdot x}}{s_y} = \frac{s_{x \cdot y}}{s_x} = \sqrt{1 - r_{xy}^2}. \tag{3}$$

The coefficient of alienation increases from zero to unity as the correlation coefficient decreases from unity to zero; that is, they vary mutually on the same scale but in opposite directions. The complement to unity of the coefficient of alienation therefore varies in the same sense as r_{xy} and may be used as an index of the prediction value of r_{xy}. Thus a "*prediction index*" may be established as

$$\text{P.I.} = 1 - \text{C.A.} = 1 - \sqrt{1 - r^2}. \tag{4}$$

It is of considerable value to study the relationship of the coefficient of alienation and of the prediction index to the coefficient of correlation. These relationships are portrayed graphically in Fig. 30, for C.A. and P.I. are continuous functions of r alone. The graph is a quadrant of a circle passing from (P.I. = 0, r = 0) to (P.I. = 1, r = 1). These limits form the only two points at which the correlation coefficient appropriately indicates prediction capacity on a correlation surface. *Thorough appreciation of the nature of this graph is essential to correct understanding of the correlation coefficient as a measure of "intensity of association" between two variables.* In our preceding discussion of the correlation coefficient we were content to accept demonstration that it passed in magnitude from zero to unity as the "intensity of correlation" moved from nothing at all to perfect rectilinear dependence. We are now in a position to interpret more fully the meaning of intermediate values.

Let us study this matter further by a question-and-answer method.

(1) *Question:* Does a correlation coefficient of 0.5 mean that half of the variation in one variable is accounted for in terms of the other variable?

Answer: No! When $r = 0.5$, slightly more than 86 per cent of the variation in the dependent variable is still unaccounted for.

(2) *Question:* At what value of r is the residual variation reduced to 50 per cent of the total?

Answer: $r = 0.86$.

(3) *Question:* How high must the correlation coefficient be before only 10 per cent or less of the variation remains unaccounted for?

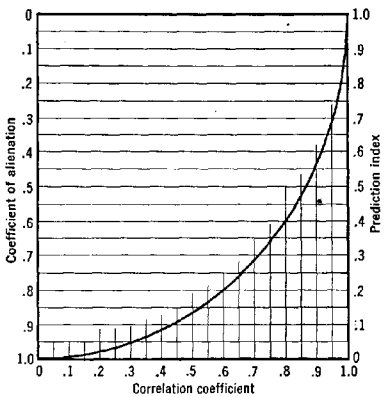
Answer: $r = 0.995$ or above.

(4) *Question:* How high must r be before 10 per cent of the variation in one variable is accounted for in terms of the other variable?

Answer: The correlation coefficient must be 0.436 or above.

FIGURE 30

FUNCTIONAL RELATIONSHIP OF THE COEFFICIENT OF ALIENATION AND THE "PREDICTION INDEX" TO THE COEFFICIENT OF CORRELATION



It is sometimes taken as a basis for condemnation of the correlation coefficient that a difference of 0.1 (or any fixed increment) between two values of r means an increasing difference in intensity of association, from the prediction point of view, as one passes from low to high values of r. This deficiency is regrettable, but no wholly satisfactory way of avoiding it by using a function of r, such as r^2 or the prediction index just considered, has as yet found general appeal. It should be understood clearly that the correlation coefficient, as a measure of intensity of association, must be interpreted with care. Spurious interpretation has its roots in lack of understanding by the interpreter and does not originate in the statistic itself.

CHAPTER 10

ERRORS OF RANDOM SAMPLING

The scientist is interested in the necessarily limited number of individual measurements with which he must deal only so far as they form a basis of generalizations concerning the much larger population or supply of such measures in a universe of discourse. In the analysis of data, the biometrician therefore faces tasks beyond those of statistically describing samples. He will be called upon to form from that information an estimate of what may be expected to be true of the supply from which the sample or samples have arisen. He will be asked to determine whether descriptive statistics such as means calculated from samples of different origins show differences of sufficient magnitude to warrant the inference that the populations which they represent are differentiated in like manner. The question arises immediately: To what extent is $\bar{x}$, or any other such statistic calculated from N individuals randomly drawn as a sample from a specified supply, trustworthy as a measure of the mean or other corresponding value of the supply?

It is a matter of common experience that descriptive statistics do vary from one random sample to another, although that variation naturally cannot be observed if only a single sample is drawn. It must be apparent that a precise knowledge of the amount and kind of such variation which may in general be expected is essential to the determination of the trustworthiness of each sample statistic as a basis of inference concerning the supply. Variation in statistics which is due to errors of random sampling must surely be susceptible of description in terms of the same quantities as are employed for defining the distribution of variates. Orderliness of variation in the latter must surely suggest the same for the former. Is not the distribution of variates also the distribution of means for samples of but one individual each?

If a sample is drawn from a set of measures, all of which are identical, it is immediately obvious that the sample will reproduce the supply precisely with respect to that measurement, regardless of the size of the sample drawn. A sample of one is sufficient to reproduce the invariant measurement. However, if another supply of measures which differ among themselves is used, it is equally obvious that a sample of one cannot reproduce the new supply. Indeed, if that supply (assumed to

be infinitely great in its total frequency) should embrace only 10 different magnitudes, one would feel it a remarkable coincidence if a sample of 10 individuals drawn at random reproduced all the supply measures. Even then the supply would not be reproduced perfectly by the sample unless in that supply all 10 measurements occurred with equal frequency. Thus the accuracy of determination of supply characteristics through analysis of a sample would seem to constitute a complex problem, apparently dependent at least on the nature of the supply variation and the size of the sample drawn.

THE RANDOM SAMPLE

Reference has been made above to samples drawn at random from some supply of measures. The concept of the random sample is so important in statistics that repeated consideration of its nature will not be amiss. From the simplest games of chance to the pinnacles of statistical inference in science it plays a crucial role. Once a supply of measures characterized by variation is completely defined, it is not a difficult matter to devise methods of drawing a representative sample in the sense that it is a random one. The proportionate occurrence of each single magnitude in that supply defines the probability of selection of that magnitude in the truly random sample. Beads, discs, or suitable objects may be drawn as a sample from a receptacle in which a very large number of them have been thoroughly mixed or "randomized"; or numbers may be assigned to individuals in the supply and a set of such numbers chosen from a table of "random sampling numbers" (such as that of Tippett [1]) to define the sample. Such samples will differ in constitution solely through "errors of random sampling," for the proportionate occurrence of each magnitude in the supply defines the probability of its being drawn in the sample.

On the other hand, when the frequency distribution of the variates in the supply itself is not known, the scientist faces a very different problem. A set of measurements arise through circumstances prescribing the material for and scope of an experiment or field of observation. The sample is readily specified, but the supply of which it is representative in the sense of being a random sample remains to be defined on logical grounds. The ease with which such logic may be corrupted by wishful thinking undoubtedly leads quite often to erroneous definition of the supply. Far too much generality is likely to be introduced into that definition, leading to conclusions which independent tests fail to substantiate.

[1] Tracts for computers, No. 15. Cambridge University Press.

STATISTICS AND PARAMETERS

In order to proceed with the discussion of this most important problem it is desirable to differentiate clearly, both in name and in symbol, between *statistics*, which describe the characteristics of samples, and the corresponding magnitudes describing the supply. Let the latter be designated as *parameters* and be symbolized where convenient by Greek letters equivalent in some way to the English letters used for the statistics. Thus σ (lower-case Greek s, called "sigma") and s may both stand for the standard deviation, but the former will refer to the value in the supply whereas the latter will pertain to the sample. For a given supply, σ is a fixed value, but s will vary from one sample to another when those samples are drawn at random from the supply. Likewise the correlation coefficient of a given normal bivariate supply may be designated by ρ (Greek "rho," lower case), the coefficient for samples being symbolized by r. Then in absence of knowledge of the parameters σ and ρ, the statistics s and r will form bases of estimation of their respective values.

Parametric designation for the mean cannot be conveniently given as a Greek equivalent to the "bar" above the symbol for the variable. One may use μ as an equivalent of m for mean. We shall, however, wish to use μ for moment coefficients of the supply. Now mean moments about zero as origin are often designated as m' to distinguish them from the moment coefficients m, for which the mean is the origin. It will be seen that

$$\frac{\Sigma x}{N} = m_1' = \bar{x},$$

$$\frac{\Sigma x^2}{N} = m_2',$$

and so forth. Omitting the subscript 1 as unnecessary and placing there the symbol for the variable, μ_x' will form a logical parametric notation for the mean of the supply of a series of x values.

From a given supply of variates x, indefinitely large in number, let samples of N individuals be drawn at random. Let it be assumed that a very large number, k, of such samples, all of the same total frequency N, has been drawn. Then k values of $\bar{x}$ will be available, and also k values of s_x. The true sampling error of each sample mean and standard deviation, so far as those statistics form estimates of the corresponding parameters, may be expressed as $\bar{x} - \mu_x'$ and $s_x - \sigma_x$ respectively. One might anticipate that the frequency distributions of the k values of $\bar{x}$ and s_x would have as more or less central values the parameters μ_x' and σ_x respectively. But what are the mean, standard deviation, and form

of the frequency distribution for each statistic? Answers to these questions may be derived algebraically, but the complexity of the derivations forbids their inclusion here. Recourse will be made at this point to the results of an actual sampling experiment which attests quite well to the known theoretical relationships.

THE SAMPLING DISTRIBUTION OF MEANS

The supply of finger-length measurements which have already been used as an illustration of a normal distribution[2] was prepared for sampling by assigning to each of the 3,000 measurements a number between 1 and 3,000. These numbers were then sampled, using the table of Tippett (*op. cit.*), samples of 4 numbers each being drawn at a time. The measurements corresponding to these numbers then formed the samples, 786 of which were so drawn. In Fig. 31 the outer histogram gives the distribution of the supply as defined by the original records. The inner histogram portrays the distribution of the 786 sample means, grouped into classes of 1-mm. range like the original records.

The curve superimposed on the distribution of means in Fig. 31 is the normal frequency distribution having the following parameters:

(1) its mean coincides with the mean of the original measurements,

$$\mu'_{\bar{x}} = \mu'_x; \tag{1}$$

(2) its standard deviation is equal to the standard deviation of the supply divided by the square root of N, the size of sample drawn,

$$\sigma_{\bar{x}} = \frac{\sigma_x}{\sqrt{N}}. \tag{2}$$

The fit of the normal curve to the results of the actual sampling test is good. This would naturally be expected if equations (1) and (2) above present the proper relationships and the form of distribution is normal. The equations may be derived algebraically and are explicitly correct. The normal form for the distribution of means may also be shown to be correct whenever the distribution of variates comprising the supply is normal.

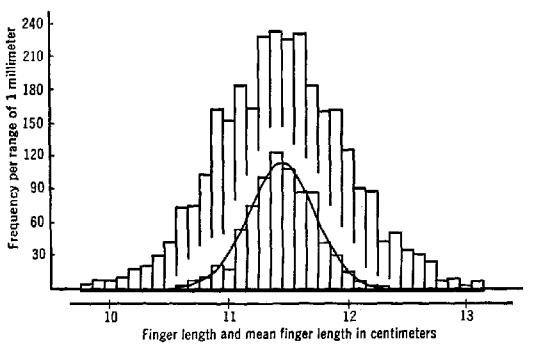
Although biological distributions in general follow I forms of curves with the normal law as a central type, they very frequently show some more or less small degree of departure from normality. How, then, are means of samples drawn at random from such supplies distributed? Equations (1) and (2) above may be derived without any assumption

[2] *Vide* page 14.

whatever as to the form of distribution in the supply. They therefore hold in *all* random sampling situations. However, it is found that the skewness and kurtosis of the supply affect the same features of the distribution of sample means. Strictly speaking, the latter distribution has one Nth part of the skewness and kurtosis of the supply curve when these features are measured in terms of the "beta coefficients." From the practical point of view in biology this pulling away of the distribution of means from the normal form is in general so small as to be of trivial consequence. Only when dealing with samples of very small frequency which have been drawn from variables with markedly anormal distribution form need there be any concern about anormality in the sampling distribution of means. With such situations we shall not deal herein; they represent special cases of comparatively rare occurrence. Let it be understood then in that which follows that N is assumed to be adequately large to render essentially normal the form of distribution of means arising from repeated random samples.

FIGURE 31

FREQUENCY DISTRIBUTIONS OF A SAMPLING SUPPLY, AND OF THE MEANS OF 786 RANDOMLY DRAWN SAMPLES OF 4 VARIATES EACH



The sampling error of any one mean, $\bar{x}$, is defined by the deviation $\bar{x} - \mu'_x$. If on the average with repeated sampling this deviation reduces to zero, then the error of sampling in means may be said to be without bias. The mean of a random sample then becomes defined as a statistic without bias. It does not tend *on the average* to overestimate or underestimate the mean of the supply from which the sample is drawn. Note that this statement does not imply that averages of random samples are

without sampling error. There remains the element of an unbiased and unpredictable error which follows the normal law having a standard deviation equal to one "root Nth" part of the standard deviation of the supply. When the supply parameters are known, then the sampling error distribution of means is known completely. If, however, the supply parameters are unknown and only a random sample from that supply is available as a basis of induction, then a very different situation faces the analyst. It is with this very practical problem that we shall now deal.

CONFIDENCE INTERVALS

One may accept the mean $\pm$ 3 standard deviations as a range within which lie practically all individuals in a normal distribution. For the sampling distribution of means, then, $\mu'_x \pm 3\sigma_{\bar{x}}$ specifies reasonable limits within which single values of $\bar{x}$ may be expected to lie, although one must remember that a very small proportion will actually fall outside even this range. If the mean μ'_x of a supply is not known, but a single value of $\bar{x}$ given by a random sample from that supply is available, then all one may infer about μ'_x is that it may have any one in a continuous range of values which would yield the given value of $\bar{x}$ through errors of random sampling. Limiting those errors to $3\sigma_{\bar{x}}$ as a working range, if $\bar{x}$ falls at the tentative limit of positive error, then

$$\mu'_x = \bar{x} - 3\sigma_{\bar{x}};$$

if the error is negative, on the other hand, and again at the provisional limit,

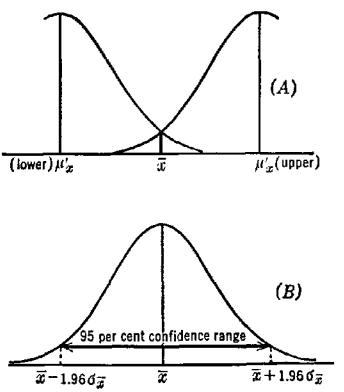
$$\mu'_x = \bar{x} + 3\sigma_{\bar{x}}.$$

It is a reasonable induction that μ'_x lies somewhere in the range specified by $\bar{x} \pm 3\sigma_{\bar{x}}$. One feels confident to a degree closely approaching certainty that μ'_x does not lie outside this range. Since any such interval is defined by sampling probability, it is customary to specify the confidence interval by the probability of sampling on which it is founded. Since $\pm 3\sigma$ in a normal distribution defines a central probability value of 99.973 per cent, the points $\bar{x} \pm 3\sigma_{\bar{x}}$ may be referred to as the 99.973 per cent confidence limits. Intervals portraying other degrees of confidence may be established in like manner. The limits specified by $\bar{x} \pm 1.96\sigma_{\bar{x}}$ define the 95 per cent confidence interval. In the upper panel of Fig. 32 the two possible values of μ'_x defined by the 95 per cent confidence limits, and the relation of the observed x to them, are given geometrically. It will readily be seen that a "confidence distribution curve" may be established about $\bar{x}$ as mean with standard deviation

equal to $\sigma_{\bar{x}}$. Such a curve is drawn in the lower panel of Fig. 32. It is of interest to note that the most likely value of μ'_x to yield the given $\bar{x}$ value is that of $\bar{x}$ itself. Contrasting with this is the equally interesting point that the confidence of μ'_x having that value is infinitesimal, as it must be for any single value without latitude.

FIGURE 32

CONFIDENCE INTERVALS WITH RESPECT TO A SUPPLY MEAN μ'_x ESTIMATED FROM A SAMPLE MEAN $\bar{x}$



(*A*) The lower and upper values of μ'_x for which the probability of $\bar{x}$ being a random sampling deviation is 5 per cent.

(*B*) The corresponding "curve of confidence" with respect to ranges within which μ'_x may lie.

The reasoning leading to the establishment of confidence limits for a parameter, as above, is straightforward and involves no assumption. The idea is fundamentally very simple and the terminology acceptable. The student should be careful, however, not to drop into the pitfall of concluding that "the probability that the mean of the supply falls in the range $\bar{x} \pm 1.96\sigma_{\bar{x}}$ is 95 per cent." That is quite wrong, for μ'_x in all cases has a fixed value. It is not a variable, and the probability attaching to any one value being correct is either one or zero according to whether that value is right or wrong. The probability used in specifying the confidence limits pertains to the *confidence* in the range embracing the parameter, not to the parameter itself lying in that range. This is a fine point, but a most important one to clear reasoning.

In the preceding discussion the unknown character of μ'_x has been observed, but no mention has been made of the unknown character of

$\sigma_{\bar{x}}$. Equational definition requires that σ_x be known before $\sigma_{\bar{x}}$ can be calculated precisely. If the mean of the supply is unknown, is it likely that the standard deviation of the supply would be known? Except in special cases the answer must be in the negative. How then may one proceed in practical situations to establish confidence limits with respect to the mean of a supply? It is at this point that some degree of approximation must inevitably be introduced if progress is to be made in providing an answer. The only basis of estimate of σ_x provided by a unique sample[3] is the variation present among the sample variates themselves. The measure of that variation is s_x, and one therefore faces the new problem of how good an estimate of σ_x is provided by s_x. The answer to this problem must be secured from study of the errors of random sampling in standard deviations.

THE ERRORS OF RANDOM SAMPLING IN STANDARD DEVIATIONS

The sampling distribution of standard deviations has been fully established by mathematical derivation only for the situation where the variates in the supply are normally distributed. It is customary to accept the relationships so given as being near enough for application in practical work with all reasonably normal biological distributions. Certainly dependable evidence to the contrary is conspicuously absent. The convention will therefore be followed herein.

Figure 33 presents the distributions of s_x for samples of size 5, 9, and 25, randomly drawn from a normal supply of standard deviation σ_x. The means of the sampling distributions are indicated by the inscribed ordinates. Two features of these distributions immediately attract attention. They are:

(1) the distributions are noticeably skew when N is small, but approach symmetrical form as N increases;

(2) the mean values of s_x are *less* than σ_x, but approach that value more closely as N increases.

With samples having a frequency of about 30 or more, no serious error is likely to be introduced in biological work if it is assumed that the distribution of standard deviations follows the normal law defined by the following equations:

$$\mu'_{s_x} = \sigma_x, \tag{3}$$

and

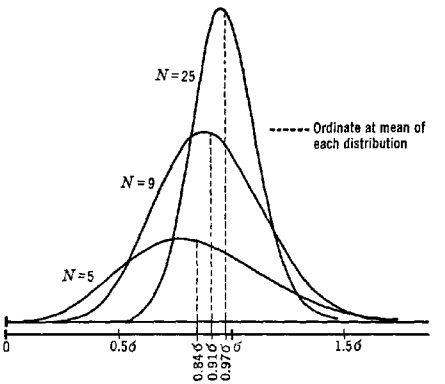
$$\sigma_{s_x} = \frac{\sigma_x}{\sqrt{2N}}. \tag{4}$$

[3] The adjective "unique" as applied to a sample in this book refers to "onlyness"; no other source of information than is provided by the sample is implied.

With sufficiently large samples it is acceptable then to conclude that the random sampling distributions of both means and standard deviations follow the normal law. In both cases the corresponding parameters form the central values, but the standard deviations of the

FIGURE 33

RANDOM SAMPLING DISTRIBUTIONS OF STANDARD DEVIATIONS FOR SAMPLES OF SIZE N DRAWN FROM A NORMAL SUPPLY OF STANDARD DEVIATION σ



sampling errors differ. It is interesting to observe from equations (2) and (4) above that the standard deviation is the more trustworthy statistic in the sense that its random sampling error is less than that of the mean. The ratio of the random sampling error in means to that in standard deviations is given on the whole by

$$\frac{\sigma_{\bar{x}}}{\sigma_{s_x}} = \frac{\sigma_x}{\sqrt{N}} \cdot \frac{\sqrt{2N}}{\sigma_x} = \frac{\sqrt{2}}{1}.$$

That is, one may have more confidence, in general, that s_x is a good estimate of σ_x than that $\bar{x}$ is a good estimate of μ'_x.

STANDARD ERRORS

The question which led to consideration of the sampling error distribution of standard deviations in the last section was one of finding from a unique sample a suitable estimate of the standard deviation of the supply from which that sample was drawn at random. Such an

estimate is needed in order that one may develop an acceptable measure of the sampling errors to which the mean itself is subject. If, in the equation

$$\sigma_{\bar{x}} = \frac{\sigma_x}{\sqrt{N}},$$

one may substitute for σ_x an appropriate estimate of it, then an estimate will be forthcoming for $\sigma_{\bar{x}}$.

With samples of large frequency there is but little latitude, relatively speaking, within which s_x may differ from σ_x. Under such circumstances it has become customary to consider s_x an adequate estimate of σ_x for substitution in the right-hand member of the equation above. The estimate so-determined of $\sigma_{\bar{x}}$ is known as the *standard error* of the sample mean.

$$\text{S.E.}_{\bar{x}} = \frac{s_x}{\sqrt{N}}. \qquad (N \text{ large.}) \tag{5}$$

The term "standard error" is one of wide application, and one may choose this occasion as suitable for emphasizing its character. It is a determinable estimate of the true but unknown standard deviation of errors of random sampling attaching to any statistic. It is formed by substituting an estimate of the needed parameter in the equation defining the true standard deviation of the statistic to which it relates. Thus, on the same grounds of reasoning, the standard error of the standard deviation of a sample is given by

$$\text{S.E.}_{s_x} = \frac{s_x}{\sqrt{2N}}. \qquad (N \text{ large.}) \tag{6}$$

This may be compared with the true but usually indeterminate formula,

$$\sigma_{s_x} = \frac{\sigma_x}{\sqrt{2N}}.$$

These estimates are really inadequate with small samples. It is important to note again from Fig. 33 that s_x does not, *on the average*, yield the supply standard deviation, σ_x, but a somewhat smaller value. From the point of view of estimating σ_x, one recognizes that s_x is a *negatively biased statistic*. When N exceeds 30, say, this bias is entirely negligible. For smaller values of N, however, it seems desirable to correct for this bias when estimating σ_x.

Reasons for this negative bias have been set forth in Chapter 4, where it was suggested that division of $\Sigma(x - \bar{x})^2$ by $N - 1$, rather than by N, represents a logical step in *direct estimation of the supply variation*. It

is most interesting to find that this logic, recognized by the astronomers since the time of Gauss, is supported by an identical correction arising from quite different reasoning. During recent decades, Fisher has vigorously urged a method of *maximum likelihood estimation* for forming "unbiased" estimates of supply parameters from the study of finite samples. This method[4] leads to the formula,

$$\text{M.L.E.}_{\sigma_x} = s_x\sqrt{\frac{N}{N-1}}$$

$$= \sqrt{\frac{\Sigma(x-\bar{x})^2}{N-1}},$$

as providing the "most likely" value of σ_x to lead to the observed standard deviation for the sample. If this estimate of σ_x is used in determining a standard error expression for the mean of a sample, one finds by simple algebra that

$$\text{S.E.}_{\bar{x}} = \frac{s_x}{\sqrt{N-1}}, \tag{7}$$

regardless of the magnitude of N. Obviously, equation (7) approaches equation (5) rapidly as N increases from very small values. When N is 30 or more the difference between the two expressions is trivial.

Proceeding then from a unique sample, one may estimate confidence limits with respect to the mean of the supply by substitution in the expression, $\bar{x} \pm k(\text{S.E.}_{\bar{x}})$, where k is the appropriate factor for the probability involved. k may be secured readily from tables of the normal probability integral for any desired degree of confidence. It must be borne in mind that, just as this estimate of the confidence interval becomes more precise as N increases, so inversely it becomes less acceptable as N decreases.

RANDOM SAMPLING DISTRIBUTION OF DIFFERENCES BETWEEN MEANS

Perhaps the most common test called for in statistical analysis is that of determining whether the difference between the means of two samples is of sufficient magnitude to justify the inference that they reflect a real difference in the means of the supplies from which those samples have been drawn. For example, in a series of over 2,000 births at a certain

[4] A simple exposition of this method for the mathematical reader may be found in the following article: W. Edwards Deming and Raymond T. Birge. On the statistical theory of errors. Reviews of Modern Physics, 6: 119-161. 1934. Reprinted with additional notes by the U. S. Dep't Agr. Graduate School, 1938.

hospital it is found that the boys are on the average a little over $\frac{1}{4}$ inch longer than the girls. Is this small difference ascribable to sampling errors, or are boys really somewhat longer on the average than girls at birth? It is common knowledge that in adult life males exceed females in stature, but the smallness of this difference at birth invites careful consideration in relation to sampling errors. This being but one of a multitude of similar problems, let us consider the general problem first.

Let N_x, $\bar{x}$, and s_x be the statistics descriptive of a first sample, and let N_y, $\bar{y}$, and s_y be those of a comparable sample of different origin. Let the supply parameters be μ_x', σ_x, and μ_y', σ_y, respectively. Then from the preceding discussion of sampling errors we know that $\bar{x}$ is merely an individual value in a normal sampling distribution of means which has its central value at μ_x'. The $\bar{y}$ value belongs in an analogous distribution centered at μ_y'. Since $\bar{x}$ and $\bar{y}$ differ in magnitude, it is desired to ascertain from our knowledge of sampling errors in means whether it is a safe inference that μ_x' and μ_y' differ in magnitude.

For purposes of discussion let us hypothecate that the means of the two supplies are identical; that is,

$$\mu_x' = \mu_y'.$$

This hypothesis must, in fact, be either true or false. If it is true, the observed difference between $\bar{x}$ and $\bar{y}$ is attributable solely to errors of random sampling. As a test of the reasonableness of the hypothesis, one naturally asks: What is the probability that a difference as great as $\bar{x} - \bar{y}$ would arise through errors of random sampling alone? A large probability would surely justify ascribing the difference to sampling effects. On the other hand, if that probability is quite small, then one must decide whether adherence to the hypothesis of no difference between the parameters is still warranted. It may, perhaps, be more logical to reject the "*null hypothesis*" in favor of its alternative, namely that

$$\mu_x' \neq \mu_y'.$$

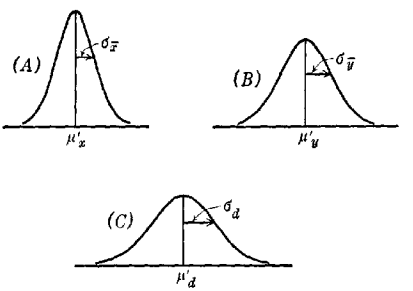
The statistical task is completed when the correct probability is secured. The interpreter must decide whether to accept or reject the hypothesis yielding the probability. The statistical investigation lays the foundation for logical processes of interpretation; it initiates, but most certainly should not be construed as terminating, a continuum in logical analysis.

The situation with respect to the distribution of differences between means arising solely through errors of random sampling is presented

geometrically in Fig. 34. Curve (A) portrays the normal distribution of values of $\bar{x}$ based on samples of size N_x, random sampling having been made from a supply of mean μ'_x. Curve (B) similarly applies to means $\bar{y}$ based on samples of size N_y drawn from a supply of mean μ'_y. If, from each of these distributions (A) and (B), one of the means they portray is drawn at random, and from the pair so drawn the difference $d = \bar{x} - \bar{y}$ is calculated, then continued repetition of this procedure will give a large series of differences of which the distribution parameters may be derived. Curve (C) portrays the curve of distribution of d, a curve of normal form for which the mean and standard deviation will now be determined by simple algebra.

FIGURE 34

DISTRIBUTION OF RANDOM SAMPLING DIFFERENCES BETWEEN ELEMENTS WHICH ARE NORMALLY DISTRIBUTED



(A) Curve of the distribution of means, $\bar{x}$
(B) Curve of the distribution of means, $\bar{y}$
(C) Curve of the distribution of differences, $d = \bar{x} - \bar{y}$

$$\mu'_d = \mu'_x - \mu'_y$$

$$\sigma_d = \sqrt{\sigma_{\bar{x}}^2 - \sigma_{\bar{y}}^2}$$

Let a very large number, k, of differences be assumed available.

Then
$$\bar{d} = \frac{\Sigma d}{k} = \frac{\Sigma(\bar{x} - \bar{y})}{k}$$

$$= \frac{\Sigma \bar{x}}{k} - \frac{\Sigma \bar{y}}{k}. \tag{8}$$

In the limit of $k \to \infty$, obviously

$$\mu'_d = \mu'_{\bar{x}} - \mu'_{\bar{y}} = \mu'_x - \mu'_y. \tag{9}$$

Now
$$s_d^2 = \frac{\Sigma d^2}{k} - \bar{d}^2$$
$$= \frac{\Sigma(\bar{x} - \bar{y})^2}{k} - \left[\frac{\Sigma(\bar{x} - \bar{y})}{k}\right]^2$$
$$= \frac{\Sigma\bar{x}^2}{k} - 2\frac{\Sigma\bar{x}\bar{y}}{k} + \frac{\Sigma\bar{y}^2}{k} - \left[\frac{\Sigma\bar{x}}{k}\right]^2 + 2\frac{\Sigma\bar{x}}{k}\frac{\Sigma\bar{y}}{k} - \left[\frac{\Sigma\bar{y}}{k}\right]^2.$$

By rearrangement of the terms in this expression,

$$s_d^2 = \frac{\Sigma\bar{x}^2}{k} - \left[\frac{\Sigma\bar{x}}{k}\right]^2 + \frac{\Sigma\bar{y}^2}{k} - \left[\frac{\Sigma\bar{y}}{k}\right]^2 - 2\left[\frac{\Sigma\bar{x}\bar{y}}{k} - \frac{\Sigma\bar{x}}{k}\frac{\Sigma\bar{y}}{k}\right]$$
$$= \quad (A) \quad + \quad (B) \quad - \quad 2(C).$$

Expression (A) is the difference between the mean square and the square of the mean of k values of $\bar{x}$. Therefore

$$(A) = s_{\bar{x}}^2.$$

Similarly,
$$(B) = s_{\bar{y}}^2.$$

Now (C) is the difference between the mean product of paired values and the product of their means. By rearrangement of the formula for the correlation coefficient between such paired values we find that

$$(C) = r_{\bar{x}\bar{y}}\, s_{\bar{x}} s_{\bar{y}}.$$

That is,
$$s_d^2 = s_{\bar{x}}^2 + s_{\bar{y}}^2 - 2r_{\bar{x}\bar{y}}\, s_{\bar{x}} s_{\bar{y}}.$$

In the limit of $k \to \infty$, these statistics become parameters of the curve of distribution of d. In that limit, however, correlation between the paired values of $\bar{x}$ and $\bar{y}$ must vanish, for they are drawn *independently* by definition. It follows then that the standard deviation of the curve of differences is

$$\left.\begin{aligned} \sigma_d &= \sqrt{\sigma_{\bar{x}}^2 + \sigma_{\bar{y}}^2} \\ &= \sqrt{\frac{\sigma_x^2}{N_x} + \frac{\sigma_y^2}{N_y}}. \end{aligned}\right\} \tag{10}$$

Derivations of more complex character show that the third-moment coefficient of the d distribution is zero, and its fourth-moment coefficient is precisely three times the squared second-moment coefficient. The curve of distribution of differences between means of independent samples therefore follows the normal law, with parameters given by equations (9) and (10) above.

The foregoing derivations have been made without any specification about identity in the means of the supplies. If it is now taken as a special case that they are identical, that is $\mu_x' = \mu_y'$, the distribution of differences between means of samples randomly drawn from those supplies will be a normal curve about *zero* as mean, with a standard deviation given by equation (10). In finding the probability that an observed difference between two means $\bar{x}$ and $\bar{y}$ is due solely to errors of random sampling, it is therefore not necessary to know what the common mean of the supplies is. Such differences are distributed normally about zero.

The probability of any given difference (or a larger one) between the means of two samples drawn at random from supplies of identical mean would be derivable precisely from tables of the normal curve functions only if the supply standard deviations were known. The relative deviate of the difference in the normal distribution defining the errors of random sampling would be given by

$$k = \frac{d - \mu_d'}{\sigma_d} = \frac{\bar{x} - \bar{y}}{\sqrt{\sigma_{\bar{x}}^2 + \sigma_{\bar{y}}^2}}. \tag{11}$$

At this point one must usually turn to estimation. Dependent on σ_x and σ_y for their values, $\sigma_{\bar{x}}$ and $\sigma_{\bar{y}}$ are not likely to be precisely determinable except in very special cases. In place of them one may choose to use the standard errors as estimates, yielding the derived equation

$$\text{S.E.}_{\bar{x}-\bar{y}} = \sqrt{\text{S.E.}_{\bar{x}}^2 + \text{S.E.}_{\bar{y}}^2}. \tag{12}$$

This expression for the standard error of the difference between two means is an estimate of the true but unknown standard deviation of the errors of sampling determining such differences. Substituting S.E._d in equation (12) for σ_d in equation (11), one has the approximation

$$k\,(=)\,\frac{(\bar{x} - \bar{y})}{\sqrt{\text{S.E.}_{\bar{x}}^2 + \text{S.E.}_{\bar{y}}^2}}, \tag{13}$$

which, of course, improves as the sample frequencies N_x and N_y increase. From k the corresponding probability may be found from the table in Appendix I.

PRACTICAL APPLICATION

The computational steps involved in testing the null hypothesis applied to the means or standard deviations of large samples are of very simple character. Analysis of the problem immediately following will exemplify this. Data secured from the archives of the Sloane Hospital,

New York City, for length of new-born infants of Irish parents yielded the following statistics:

Sex	Number	Mean	St. Dev.
Male (x)	1136	51.96 cm.	2.181 cm.
Female (y)	1071	51.22 cm.	2.189 cm.

These data may be analyzed to determine whether the sex differentiation with respect to mean value and variability shown by the samples justifies the inference that Irish male offspring are in general longer and less variable in length than females at birth. Since the standard deviations of the supplies are unknown, one must depend on standard errors, which for such large samples will provide excellent estimates.

1. The difference in means

$$d = (\bar{x} - \bar{y}) = 0.74 \text{ cm.}$$

$$\text{S.E.}_{\bar{x}}^2 = \frac{2.181^2}{1136} = 0.004187.$$

$$\text{S.E.}_{\bar{y}}^2 = \frac{2.189^2}{1071} = 0.004474.$$

$$\text{S.E.}_d = \sqrt{0.004187 + 0.004474} = 0.093.$$

$$k\,(=)\,\frac{d}{\text{S.E.}_d} = \frac{0.74}{0.093} = 7.96.$$

This value of k is beyond the range of the table in Appendix I. The probability of such a difference being due solely to errors of sampling is in fact less than one in a million. There can be no doubt then that the observed difference between the sample means has only a most remote chance of arising solely from errors of random sampling. A real difference of like sign in the supply means is therefore indicated. Using the 95 per cent confidence limits, one would estimate that, while 0.74 cm. is the most likely difference between the supply means, the difference may well be as low as

$$0.74 - 1.96\,\text{S.E.}_d = 0.56 \text{ cm.},$$

or as high as

$$0.74 + 1.96\,\text{S.E.}_d = 0.92 \text{ cm.}$$

The smallness of the difference (most probably less than 1 cm.) makes this problem of academic rather than practical interest, of course.

2. The difference in variability

With N as large as it is in each sample, the errors of sampling in the standard deviations must follow the normal law very closely. Therefore the difference between the standard deviations may be analyzed precisely as in the case of the means.

$$d = (s_y - s_x) = 0.008 \text{ cm.}$$

$$\text{S.E.}^2_{s_y} = \frac{2.189^2}{2(1071)} = 0.002237.$$

$$\text{S.E.}^2_{s_x} = \frac{2.181^2}{2(1136)} = 0.002093.$$

$$\text{S.E.}_d = \sqrt{0.002237 + 0.002093} = 0.066.$$

$$k(=) \frac{0.008}{0.066} = 0.12.$$

From Appendix I, one finds that the probability of a larger difference than this arising solely through sampling errors is greater than 90 per cent. It would therefore seem most foolish to ascribe any significance to this particular difference. Indeed, wide experience with such comparisons of variability in length at birth is convincing that boys are somewhat *more* variable than girls. In the preceding example it appears that a sampling error has probably more than offset a real difference of opposite sign.

A second example may be given in terms of samples of rather small size. Two chemists made 20 determinations each of the protein content of a sample of flour sent to them for analysis. The determinations, with tests of the differentiation in mean and standard deviation of each set, are given in Table 20. Again there appears rather conclusive evidence that the difference in the mean values is not due solely to errors of sampling. Analyst x has recorded almost 0.1 per cent more protein than analyst y. Evidence that one is any more consistent in his analyses than the other is lacking; the standard deviations certainly do not differ significantly.

One may well feel somewhat insecure about the inference that the difference in means is significant, because standard errors derived from rather small samples are likely to be poor substitutes for the true standard deviations of random errors. The point may be tested further by aiming to set up a theoretical value for that standard deviation as follows. If σ is the true error of analysis for these two workers, how large

would it need to be to cast doubt on the significance of the difference between the means? An answer to this question will require, as a prerequisite, selection of some definite "level of significance." For demonstration purposes let us choose the 5 per cent level. Then the observed

TABLE 20

REPLICATE DETERMINATIONS OF PROTEIN CONTENT OF THE SAME FLOUR SAMPLE BY TWO ANALYSTS, WITH STATISTICAL TESTS FOR SIGNIFICANCE OF DIFFERENTIATION IN MEAN AND STANDARD DEVIATION

x	y	Tests of significance
9.95	9.94	$\bar{x} = 10.0495.$
10.01	9.91	$s_x = 0.0397.$
10.04	10.01	
10.01	9.94	
10.09	9.96	$\bar{y} = 9.9525.$
10.06	9.98	$s_y = 0.0455.$
10.11	9.96	
10.05	10.03	
10.07	9.94	$\bar{x}-\bar{y} = 0.0970.$
10.08	9.91	S.E. $\bar{x}-\bar{y} = 0.0135.$
10.03	9.99	$k\ (=)\ 7.19.$
9.97	9.99	$P_k < 10^{-6}.$
10.07	9.93	
10.05	9.93	
10.04	9.86	$s_x-s_y = 0.0058.$
10.06	9.86	S.E. $s_x-s_y = 0.0095.$
10.06	10.03	$k\ (=)\ 0.61.$
10.07	9.96	$P_k > 0.5.$
10.11	9.96	
10.06	9.96	

difference in means, 0.0970, must be 1.96 times σ_d to be on this level, making σ_d equal to 0.05 approximately. Now

$$\sigma_d = \sqrt{\frac{\sigma_x^2}{N_x} + \frac{\sigma_y^2}{N_y}},$$

and if σ_x equals σ_y, then with N equal to 20 in each case, it follows that

$$0.05^2 = \frac{\sigma_x^2}{20} + \frac{\sigma_y^2}{20} = \frac{\sigma^2}{10},$$

or $$\sigma^2 = 0.025,$$

and $$\sigma = 0.16 \text{ approx.},$$

where σ is the random error of analysis common to both chemists. One may now ask: Do the observed values of s_x and s_y appear consistent with this value of σ? If σ equals 0.16, then, for samples of 20, σ_s would equal 0.025, and the values of s_x and s_y would most probably exceed $0.16 - 2\ (0.025) = 0.11$. But s_x and s_y are *both* far less than this value. Therefore it seems that the analysts differ quite significantly in the mean values they report. What then is the true protein content of the flour sample? It may be well at this point to leave the reader to his own reflections.

The data presented as Table 5*a* in Chapter 2 provide opportunity for the reader to apply tests of significant differentiation in means and in standard deviations. The samples of brain weights for the two sexes fulfill the condition of being independent of each other, as is required by the above reasoning. This condition is not fulfilled by the data on statures of husbands and wives given as Fig. 18, wherein paired values arise and significant correlation has been demonstrated. We will now consider the problem of testing the significance of the mean difference between correlated pairs of variates.

PAIRED VARIATES

Every test of significance of the difference between the means of two samples is motivated, to some extent, by a desire to expose or refute the influence of some factor, or constellation of factors, which may be deemed to have exerted an influence of a "causative" nature. An investigation of the influence of such factors will be well designed when all factors other than those under consideration are precluded from influencing the results of the research. One of the best ways to achieve this end is to pair the variates in two series for all influential factors other than those under scrutiny. The influence of a single nutritive factor on growth of experimental animals, for instance, may be investigated through adequate design in this manner. Experimental rats might be paired for genetic constitution, age, and weight at the beginning of the experiment, one member going to diet group A and the other to diet group B. Assuming that during the experiment the rats are subjected to the same environmental conditions, and ingest equivalent quantities of the two diets, any significant difference in their mean growth responses may be ascribed to the dietary factor. Any such difference is, at least, not influenced by the factors for which initial pairing was made.

A direct approach in testing the significance of any difference in response of the groups is to secure the mean difference between the

pairs, and ascertain the probability that it might arise by chance. The equation,

$$k\,(=)\,\frac{\bar{d}}{\text{S.E.}_{\bar{d}}}, \tag{14}$$

may readily be solved for k when it is recognized that

$$\text{S.E.}_{\bar{d}} = \frac{s_d}{\sqrt{N-1}},$$

where N is the number of differences, or pairs of rats.

It may, however, be desired to correlate the pairs for some purpose. In that event the individual differences between pairs need not be secured. If x and y designate the variates for the two groups, then

$$\bar{d} = \bar{x} - \bar{y},$$

and

$$s_d = \sqrt{s_x^2 + s_y^2 - 2r_{xy}s_x s_y}. \tag{15}$$

The reader may easily verify these equations for himself by following the pattern given in the closely analogous derivation of equations (8) and (10) on page 140 in this chapter. It follows directly, also, that

$$\text{S.E.}_{\bar{d}} = \sqrt{\text{S.E.}_{\bar{x}}^2 + \text{S.E.}_{\bar{y}}^2 - 2r_{xy}\,\text{S.E.}_{\bar{x}}\,\text{S.E.}_{\bar{y}}}. \tag{16}$$

The importance of the factor r_{xy} in equations (15) and (16) may readily be appreciated from consideration of the following problem, if not on purely theoretical grounds.

In 1925 the Minnesota State Testing Mill studied 56 carlots of wheat. An official dockage assessment on each carlot was made by the State Grain Inspection Department, and another test was made independently at the mill laboratory. The basic data need not be given here,[5] as we need only certain derived statistics.

x = official test of dockage, in per cent
y = mill test of dockage, in per cent
$N = 56$
$\bar{x} = 3.39$ $\qquad$ $\bar{y} = 3.94$
$s_x = 2.29$ $\qquad$ $s_y = 2.26$
$r_{xy} = +\,0.97$

When the correlation is overlooked, equation (12) being used then in error, one finds $k\,(=)\,1.27$, indicating an insignificant difference in the two series of dockage tests. However, the correlation must properly be recognized by using equation (16). Then one finds that $k\,(=)\,7.03$, indicating a highly significant discrepancy between the dockage evaluations of the two laboratories.

[5] A fuller discussion may be found in J. Am. Soc. Agron., 23: 558–571. 1931.

This is but one of a multitude of examples which might be given to illustrate how truth may be lost in a significance test involving paired determinations when the correlation between the variables is overlooked. The reader will readily discern that if a comparison of the two laboratories' determinations had been made on approximately the same numbers of *different* carlots, both series being random samples from arriving shipments, the difference in means would probably have been insignificant. Tremendously greater sensitivity of the investigation is achieved in this case by confining the analyses to the same series of materials, or using the principle of paired observations.

SIGNIFICANT DIFFERENTIATION

The terms "significant difference" and "insignificant difference" are frequently used with such finality of judgment in statistical writing that it is not surprising to find many students of science who aspire to command the means of such conclusive proof. They naturally essay to learn how to make statistical tests. To do so without appreciating fully the reasons why the test is appropriate is to court trouble. This is particularly true if tests of significance are undertaken when only the simple steps of computation are known. It is particularly desirable, then, that careful consideration be given to the implication of the term "significant difference," which arises directly from interpretation of the probability that the difference between two statistical values being compared would arise through errors of random sampling alone.

The foregoing data on sex differentiation in length at birth showed that, on the average for the samples considered, boys exceeded girls in length by about $\frac{3}{4}$ cm., or slightly over $\frac{1}{4}$ inch. In the light of purely practical considerations this must surely seem an unimportant difference. And yet one finds by statistical analysis that the observed difference or a larger one would not arise through errors of sampling alone more than once in a million times. Does this first pair of samples happen to be one of those exceedingly rare ones, or is it more reasonable to conclude that boys *are* longer on the average than girls at birth? It would seem most foolish to cling to the null hypothesis in view of the evidence, and so the interpreter naturally rejects it in favor of the alternative conclusion that there is a real differentiation of the sexes with respect to length at birth.

The difference provided by the samples is taken to *signify* a difference of like character in the supplies of male and female infants from which the samples were drawn, and is called a *significant difference*. When the probability of the difference arising through sampling errors alone is large, as in the comparison of the standard deviations for length at birth,

then significance is not ascribed to the difference and the designation "insignificant" is used. Thus the verdict hinges entirely on the magnitude of the probability yielded by application of the null hypothesis.

It would appear, then, that at some point on the probability scale there is a dividing line segregating the "significant" from the "insignificant" probability. Such a conclusion, however, is merely a half truth. Two other most important factors must be considered. The transition from one judgment to the opposite one as one advances along the probability scale may logically take place gradually over a range of the scale instead of sharply at a point. Also, the consequences of rejecting one hypothesis in favor of the other must be considered in location of the transition zone (or zone of doubt), be it wide or narrow.

ERRONEOUS INFERENCE

The parameters being considered in the null hypothesis must be, in fact, either identical or different. From study of the available data given by the samples the statistician wishes to infer the correct statement from these two alternatives. That is, his inference when made will be either right or wrong. If he should draw the wrong inference, he must naturally face responsibility for the practical consequences of his mistake. Such erroneous inferences must fall into one of two classes:

(A) insignificance is claimed when a real difference in the parameters exists, or

(B) significance is claimed where none exists.

Errors of these two classes have entirely different consequences.

Erroneous inferences falling under class (A) above will cause the rejection (complete or tentative) of the data as being insufficient to warrant the claim that there is a real difference between the parameters being considered. This action essentially demands that more evidence is required before significance may properly be inferred. That evidence may, of course, be sought through more refined or extensive investigation if the problem is important enough to warrant continuance of the research.

Errors of class (B), on the other hand, ordinarily terminate further research in verification. Indeed, acceptance of a verdict of significance may well initiate a developmental program of new investigations. Such a program will be based on a wrong premise, of course, if the initial conclusion of significance was faulty. Therefore errors of class (B), assigning significance where none really exists, are likely to be the more serious by far in their immediate effect on the progress of science. The pro-

cedure of inference from a probability derived under the null hypothesis should, then, be characterized by defenses against making an error of class (*B*) above.

The probability of differences between sample statistics being due to errors of random sampling may, of course, take any value from unity down to zero. Between the values of 1 and 0.5, the probability is so high as to cause acceptance of the null hypothesis without question, in the absence, of course, of other relevant evidence. As the probability falls below 0.5, however, the evidence may be taken to favor increasingly the rejection of the null hypothesis. As a defensive measure against making an error of class (*B*) above, it is customary among statisticians to require that the probability be less than 0.05 before significance is claimed, that is, before a real difference in parameters is inferred and the null hypothesis is rejected. Any such level of probability beyond which significance is claimed is known as a *level of significance.*

Let it be assumed now that a very large number of tests of significance are to be made by a statistician who chooses to depend on a 5 per cent level of significance. Let it further be assumed that, unknown to him, the differences are all actually due to random sampling errors and are unselected for magnitude; the pairs of samples may be assumed deliberately drawn from the same supply, but this is really immaterial. In applying the test in each case arising from the null hypothesis, he will find that 5 per cent of the differences actually exceed the chosen "level of significance." He would normally take these to signify real differences in the parameters of reference, if his suspicion to the contrary was not aroused. But actually these 5 per cent of "significant differences" are by definition only random sampling differences from supplies with identical parameters. That is, *under the specified conditions* the inferences will be wrong 5 times in 100 when the 5 per cent level of significance is used, and the percentage error will be given always by the level of significance used.

It is most important to recognize in the situation just discussed that all differences were postulated as being due solely to errors of random sampling. Only errors of class (*B*) above were possible. Such a condition is not likely to prevail over any substantial length of time in statistical experience. Let us now reconsider the problem modified to parallel practical situations. Let us mix with the above paired samples, differing solely by sampling errors, a number of pairs of samples for which the parameters are really different. In testing the null hypothesis, both types of error may now be made, but let us focus attention on the more serious ones comprising class (*B*) only. The proportion of wrong inferences claiming significance when none exists will now be

diluted to something below 5 per cent because many of the claims of significance will now be right, and these correct ones must form more than 5 per cent of all really different paired samples. Thus a 5 per cent level of significance adopted in practical statistical work will lead to a *maximum* error of 1 in 20 in claiming significance.

If this proportion of class (*B*) error is too high, then it is a simple matter to shift the critical value for inferring significance to some lower level, such as 1 per cent, or even less. Such a shift must, of course, lead to a corresponding increase in errors of class (*A*). The price of avoidance of errors of one class in establishing a level of significance is inevitably one of committing more errors of the other class.

The problem of choosing an acceptable level of significance is to find an appropriate balance between two types of error of entirely different character. This balance must logically be subject to change from one situation to another. The fundamental problem is of the relative importance of the two types of error to the particular investigation involved. Unless the relative importance is constant through all problems, it is obviously not reasonable to establish a critical value of probability for all situations as bases of statements of significance. It might be an acceptable risk to endanger the life of 1 experimental rat in 20, but the same attitude would hardly be sustained if human life is involved. It is on the grounds of such reasoning that the writer of this discussion feels that judgments of significance should not be reached purely on the relation of a determined probability to some fixed critical value.

The research worker should recognize clearly the nature of this probability. It does *not* measure the probability that the observed difference between the statistics did arise solely through sampling errors. It merely defines the probability of a difference as great as the one observed arising *if* there is no difference between the supply parameters. On the basis of this evidence he is to decide whether the null hypothesis should, for purposes of scientific inference, be rejected. His decision should properly be guided by the practical importance of the consequences engendered if his conclusion should prove to be the wrong one. More convincing evidence, such as is given by a lower probability, may well be desired in some cases than in others.

CHAPTER 11

SAMPLING ERRORS OF THE CORRELATION COEFFICIENT

The coefficient of correlation was introduced by Galton and developed by Pearson as a statistic descriptive in a relative manner of the intensity of association between variables adhering to the normal law of bivariate frequency distribution. It is widely used in biological investigations today as providing satisfactory fulfillment of this descriptive function. In other cases, proper attention to the fact that its usefulness in this connection is restricted to normal surfaces is often lacking. The existence of rectilinear regressions alone is not adequate for the logical comparison of correlation coefficients. Lack of regard for this point (and others) has led to many erroneous interpretations, criticisms appropriate to the interpreter then being directed against the statistic. Undoubtedly a large proportion of biological investigations stop at correlation coefficients when they should properly be carried forward to regression equations and studies of residual variations, for definition of which the correlation coefficient may serve as an excellent generating function. There remain, nevertheless, many situations wherein direct comparison of correlation coefficients is fully appropriate to solution of the problems under analysis.

Two important questions arise in consideration of the magnitudes of correlation coefficients calculated for normal surfaces as measures of intensity of association. They are:

(1) How large should the value of the correlation coefficient be, when based on N paired observations, to justify the conclusion that correlation really exists in the supply?

(2) When is the difference between two correlation coefficients sufficiently large to warrant assumption that the supplies from which the samples were drawn really differ in their intensities of association?

Both these questions flow immediately from recognition of the fact that the correlation coefficient, like other statistics, must be expected to be influenced in its magnitude by errors of random sampling.

Let the Greek symbol ρ designate the correlation coefficient for an infinitely large supply of paired values following the normal law. From this supply let samples of size N be considered drawn at random. The correlation coefficients r of these samples may be anticipated to be dis-

tributed about ρ as a more or less central value. Of what form is this sampling distribution of r, and how may one determine from it the probabilities appropriate as bases for answering the two types of questions above?

THE SAMPLING DISTRIBUTION OF r WHEN ρ EQUALS ZERO

Having no other information about correlation in the supply than is given by the sample itself, one naturally faces the possibility that ρ is zero and that the magnitude of r, however large it be, might really be fortuitous. What is the sampling distribution of r when there is no correlation in the supply yielding the samples? If, from a normal supply of paired but uncorrelated measures, samples of a given size are considered drawn at random, one would not anticipate that the correlation coefficients calculated for these samples would all equal zero, nor indeed that any one would equal zero exactly, except by chance. However, a clustering about zero might logically be expected for the values of r; a frequency distribution of symmetrical form with zero as its mean. One finds it unreasonable to construe any factor consistent with random sampling that would give other than a symmetrical distribution of r around the parametric value of zero. Whatever chance factors arise to yield samples with a certain positive value of r must surely be expected to lead in the long run to an equal number of samples with correlation coefficients of the same numerical magnitude but of opposite sign. The normal bivariate frequency distribution with ρ equal to zero is radially symmetrical about its center. Also, the tendency for all but the smallest samples at least must be to reproduce the supply characteristic of zero correlation. A symmetrical I-shaped curve seems inescapable under such conditions. However, a strictly normal sampling distribution of r cannot be realized technically in this situation because a normal curve about zero as center must be of infinite range, whereas the scale of r is confined to the range between -1 and $+1$. What precisely is the form of this sampling distribution of r when ρ is equal to zero?

It is idle to attempt mathematical derivation of the sampling distribution of r with no more than simple algebra as a tool. One may more profitably turn here to geometric representation of the theoretical solution of the problem, first reached on partly intuitive grounds by Student in 1908.[1] In Fig. 35 the precise sampling distribution curves of r when ρ equals zero are drawn in full line for three values of N. Two features will be noted immediately from this figure:

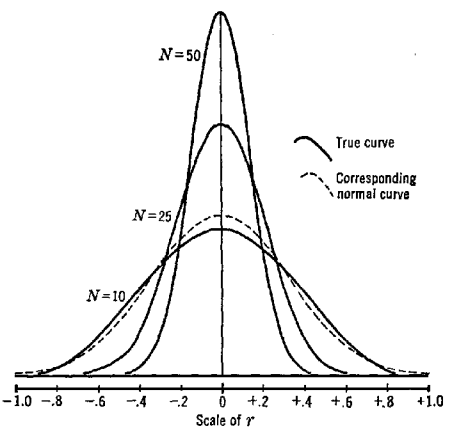
[1] Student. Probable error of a correlation coefficient. Biometrika, 6: 302-310 1908.

(1) for small samples, r may frequently take quite high values solely through errors of sampling;

(2) as N increases, the probability of r exceeding any fixed value falls rapidly.

FIGURE 35

RANDOM SAMPLING DISTRIBUTIONS OF r, BASED ON SAMPLES OF SIZE N, WHEN ρ EQUALS ZERO



These distribution curves are actually so close to the normal form that the difference is essentially negligible except for rather small values of N. The broken-line curve in Fig. 35 is the normal distribution having the same mean and standard deviation as the true curve for N equal to 10. Even there the discrepancy between the two curves is not very great. For N equal to 20 and above, the lack of concordance between the true curve and the corresponding normal curve is so small as to be of no practical consequence. It may be noted that the kurtosis of the true curve is negative. It is "flatter topped" than the normal curve of like mean and standard deviation. While the range for all the true curves is theoretically -1 to $+1$, the frequency beyond three standard deviations may for practical purposes be ignored, at least when N equals or exceeds 20.

The normal curve may well be accepted to portray with adequate approximation the sampling distribution of correlation coefficients when ρ is equal to zero and N is 20 or above. In order to define this closely approximating normal curve it is necessary, of course, to know the stand-

ard deviation of the true distribution. Student derived the definitive equation as

$$\sigma_r = \frac{1}{\sqrt{N-1}}. \tag{1}$$

Biologists confining attention to large samples frequently ignore the -1 in the denominator, but one may just as well choose to retain it to avoid changing the formula when N does become rather small.

One may proceed on the basis of this information to solve the very practical problem of determining whether any given value of r, calculated for a sample of size N equal to 20 or above, may be accepted as signifying correlation in the supply from which the sample is drawn. The procedure is very simple. The null hypothesis is that ρ actually is zero. One must then determine the probability that the given r is a deviation from zero due to errors of sampling. Using equation (1) above, the probability of any value of r or a greater one arising through errors of random sampling will be given with adequate accuracy for most practical purposes by calculating the relative deviate k of r in the normal distribution of zero mean, and then referring to tables of the normal probability integral. Now

$$k_r = \frac{r-0}{\frac{1}{\sqrt{N-1}}}$$

$$= r\sqrt{N-1}.$$

Remembering that when k equals or exceeds 2 the probability of the deviate being exceeded is 0.045 or less, one may perhaps choose to formulate the guiding rule that, when $r\sqrt{N-1}$ does not exceed 2, the correlation coefficient of the sample does not signify real correlation in the supply. This rule, of course, adheres closely to acceptance of the 5 per cent level of significance as the demarkation point of significant deviations.

When the probability derived under the null hypothesis proves to be sufficiently small, the assumption that ρ is equal to zero may logically be rejected in favor of its alternative, namely, that correlation does exist in the supply. What the value of ρ is precisely in the latter event, one cannot say. One may choose to accept the observed r as an estimate of the value of ρ, knowing, of course, that rejection of the null hypothesis is tantamount to acceptance with confidence that ρ is of the same sign as r.

The example of the normal bivariate surface used as the first illustration in Chapter 7 yields a correlation coefficient of $+0.28$ for assortative

mating for stature in 1,079 English couples. Since $\sqrt{N-1}$ *equals 32.8,* k_r equals 9.2. There can be no doubt of the statistical significance of this correlation. Had the same correlation coefficient ($r = +0.28$) been found for a series of but 50 couples, would it still be significant? This time $\sqrt{N-1}$ equals 7 and k_r becomes 1.96, a value precisely on the 5 per cent level of significance. In the absence of any other information, one might be hesitant in claiming significance in such a case. Once in 20 times with samples of 50, r would exceed $+0.28$ solely through random sampling errors. In such a situation more data might be sought before a decision was reached.

It should be recognized clearly that a correlation coefficient designated as statistically insignificant does not prove lack of association in the supply. It merely indicates, in the absence of other evidence to the contrary, that the assumption of absence of correlation in the supply is sufficiently consistent with the observed value of r to warrant its serious consideration as an appropriate hypothesis.

THE SAMPLING DISTRIBUTION OF r WHEN ρ IS NOT ZERO

It has been shown in the preceding section that the errors of sampling to which r is subject when ρ is actually zero are very great when the size of sample approaches small magnitudes. With 12 or fewer cases per sample, r is found to exceed the value 0.5 in approximately 10 per cent or more of cases, solely through errors of random sampling. By way of contrast one may turn to the question of the sampling error in r when ρ takes the limiting value of unity. Under this condition all points in the supply distribution fall on a straight line and therefore every sample must reflect the same situation. The sample coefficient will always be unity when ρ is unity, errors of sampling having vanished entirely, regardless of the size of the sample.

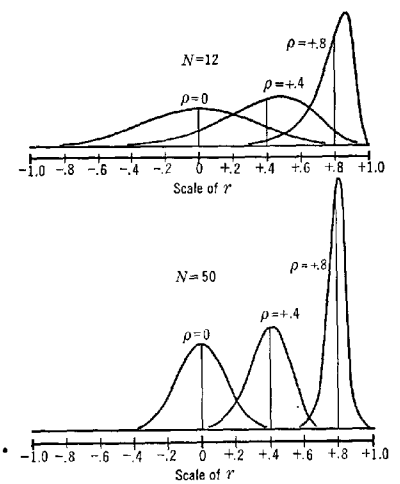
The sampling distribution of the correlation coefficient reduces to relatively simple situations when ρ takes its limiting values of zero and unity. On the basis of this knowledge of limiting conditions one may be tempted to anticipate that the sampling distribution of r might follow the simple transition of a reasonably normal curve, collapsing in standard deviation from $\frac{1}{\sqrt{N-1}}$ to zero, as ρ is moved from zero to the limiting value of unity. While such expectation is verified with respect to the magnitude of the standard deviation of errors of sampling, it is far from true with respect to the form of the distribution.

The nature of the sampling distribution of r for values of ρ other than

zero was not established precisely until 1915, when Fisher[2] derived the equation without approximation. On the basis of calculations by Soper *et al.*,[3] facilitated by this work, the sampling distribution curves of r for three values of ρ and two sizes of sample are given in Fig. 36. The upper panel relates to samples of but 12 individuals each, the lower panel giving the corresponding curves when N is 50, ρ being taken as zero, 0.4, and 0.8 in both panels.

FIGURE 36

RANDOM SAMPLING DISTRIBUTIONS OF r FOR SPECIFIED VALUES OF ρ AND N



The purpose in presenting Fig. 36 is to give some indication of the opposing influences exerted on the form of the sampling distribution of r by increasing the magnitudes of ρ and N. Starting from a symmetrical and fairly normal curve when supply correlation is absent, the form of distribution becomes increasingly skew as ρ is increased but N is held constant. For any given value of ρ, on the other hand, as N is increased the sampling distribution of r has its skewness diminished and approaches

[2] R. A. Fisher. Frequency distribution of the values of the correlation coefficient in samples from an indefinitely large population. Biometrika, 10: 507–521. 1915.

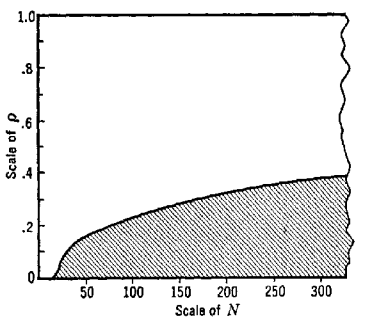
[3] H. E. Soper, A. W. Young, B. M. Cave, A. Lee, and Karl Pearson. On the distribution of the correlation coefficient in small samples. Biometrika, 11: 328–413. 1917.

the normal form. It is possible then to establish on a (ρ, N) coordinate surface an area within which the normal curve may serve fairly well as a representation of the sampling distribution of the correlation coefficient. The shaded area in Fig. 37 demarks such a zone within which the normal curve might be used with adequate approximation, its mean and standard deviation being given by

$$\left.\begin{aligned} \mu_r' &= \rho, \\ \text{and} \qquad \sigma_r &= \frac{1-\rho^2}{\sqrt{N-1}}. \end{aligned}\right\} \tag{2}$$

FIGURE 37

DEMARKATION OF AN EMPIRICALLY DETERMINED ZONE (SHADED) ON THE SURFACE OF (ρ, N), WITHIN WHICH THE SAMPLING DISTRIBUTION OF r MAY BE CONSIDERED PRACTICALLY NORMAL



The chief interest of Fig. 37 lies not in its demarkation of a zone of (ρ, N) within which a normal curve approximation might be satisfactory for the sampling distribution of r, but rather in its indication of the wide range of values of ρ and N within which normal curve approximations will *not* be adequate. In the memoir just cited, Fisher (1915) drew attention to the feature of rapidly intensifying skewness of the sampling distribution of r as ρ approaches its upper limit of unity, indicating the unsuitability of skeleton tables of the probability integral of the true distribution in that region. In 1921 he[4] removed all these difficulties by

[4] R. A. Fisher. On the "probable error" of a coefficient of correlation deduced from a small sample. Metron, 1 (No. 4): 1–32. 1921.

an ingenious transformation of scale which converted *all* the sampling distributions of r into reasonably normal curves. The transformation[5] is

$$z_r = \tfrac{1}{2}[\log_e (1 + r) - \log_e (1 - r)]$$

$$= 1.1513 \left[\log_{10} \frac{1 + r}{1 - r}\right]. \tag{3}$$

The principle of rectilinear transformation of scale has been stressed in these pages as providing a method of coding leading to the utmost attainable simplicity in calculation. In like manner Fisher's curvilinear transformation of the r scale serves to reduce a continuous array of rapidly changing curve forms to an almost invariant form, approaching the normal law so very closely that the latter provides an excellent approximation for all but very small samples.

This contribution of Fisher is notable for its reduction of a very complex problem to one of great simplicity. Probabilities corresponding to values of z_r may be obtained with considerable precision from tables of the normal curve. These probabilities, however, must be identical with those pertaining to the corresponding values of r itself. The frequency of occurrence of r is not altered by transformation; the scale only is transformed. If ζ is the transformed value of ρ following equation (3), then Fisher has shown the mean and standard deviation of the z_r curve to be

$$\left.\begin{aligned} \mu_z' &= \zeta + \frac{\rho}{2(N-1)}, \\ \text{and} \qquad \sigma_z &= \frac{1}{\sqrt{N-3}}. \end{aligned}\right\} \tag{4}$$

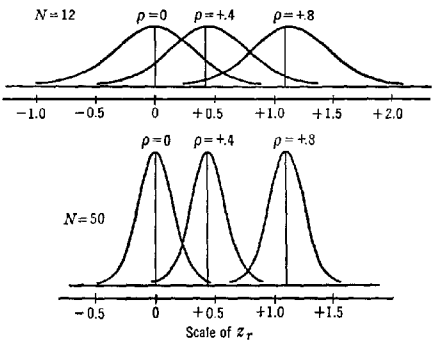
The z_r curves secured by transformation of the six different sampling

[5] It may be of interest to the mathematical reader to note that the dimensional ratio of the major to the minor axis for the normal surface ellipse, plotted in terms of relative deviates, is $\sqrt{\frac{1+r}{1-r}}$. This ratio is, of course, very intimately related to intensity of association. The transformation suggested by Fisher is simply one of replacing r by the natural logarithm of this ratio. It is readily seen from Fig. 22 that, as ρ passes from zero to unity, this ratio changes from unity to "infinity." Therefore the logarithm of the ratio varies over the range from "minus infinity" to "plus infinity."

distributions of r in Fig. 36 are given in Fig. 38. The constancy of these curves in form and standard deviation for each value of N will be noted immediately; the contrast with the parent curves in Fig. 36 is striking. The probability that any given deviation of r from ρ, or of one value of r from another, would arise through errors of random sampling may be determined by transformation to z_r, proceeding with the analysis as already detailed for normally distributed statistics.

FIGURE 38

z_r TRANSFORMATIONS OF THE RANDOM SAMPLING DISTRIBUTIONS OF r GIVEN IN FIGURE 36



The tables of Appendix II to this volume have been prepared to expedite the initial transformation of r into z_r. Simple arithmetic interpolation between the tabled values of z_r will prove adequate for determination of the values not actually given. The application of the transformations may perhaps be dealt with more advantageously now through consideration of specific numerical cases.

Problem 1. What are the values of z_r corresponding to correlation coefficients of +0.6125 and +0.9015?

From Appendix II, Table A, we have

r	z_r	Increment in z_r
0.61	0.7089	
		0.0161
0.62	0.7250	

When $r = 0.6125$, then

$$z_r = +0.7089 + 0.25\,(0.0161) = +0.7129.$$

From Appendix II, Table B, we have

r	z_r	Increment in z_r
0.901	1.4775	
		0.0053
0.902	1.4828	

When $r = +0.9015$, then

$$z_r = +1.4775 + 0.5(0.0053) = 1.4802.$$

Problem 2. Is a correlation coefficient of $+0.6125$, based on only 12 pairs of measurements from a normal supply, significantly less than the value of $+0.9015$ found for a comparable series of 19 pairs of measurements?

Coefficient	z_r	N	$\sigma_z = \frac{1}{\sqrt{N-3}}$
For $r_1 = +0.9015$	$+1.4802$	19	0.25
For $r_2 = +0.6125$	$+0.7129$	12	0.33
Difference	0.7673		

Since the standard deviation of z_r is independent of ρ, the above values involve no approximation. They are not standard errors but true standard deviations of normal sampling-error curves, of which the means are not known because ρ is in each case unknown. The problem is to test whether both samples may have come from supplies having the same value of ρ, the analysis logically reducing in many problems to testing whether both samples may have come from the same supply. This is a matter again of examining the null hypothesis. If ρ_1 is the same as ρ_2, then the two sampling distributions of z_r above have the same mean value. It follows directly that the mean of the distribution of differences arising solely through errors of random sampling will be zero. Also, the random differences curve will be normal in form with standard deviation,

$$\sigma_d = \sqrt{0.33^2 + 0.25^2} = 0.4167.$$

The situation is precisely analogous to that in testing the significance of the difference between normally distributed means. Securing the relative deviate of the observed difference in this distribution,

$$k = \frac{0.7673}{0.4167} = 1.84,$$

and from Appendix I,

$$P = 0.066.$$

Thus the observed difference between the two sample correlation coefficients would arise 6 or 7 times in 100 solely through errors of random sampling. May one appropriately conclude in this case that there is insufficient evidence to warrant the assumption of a significant difference between the two correlation coefficients? Since P is less than 0.5, the evidence favors differentiation, but the margin of doubt is not negligible by any means. Not knowing to what biological relationships these coefficients pertain, one is unable to consider the issue of the practical importance of inferring real differentiation erroneously.

Small samples have intentionally been chosen for the above problem for two reasons. First, the technique itself is probably accurate enough

for practical purposes even with N as small as 10. Second, the problem serves to indicate the relatively large differences in correlation coefficients which may arise through errors of random selection with samples of small size, even in the upper reaches of the scale where increments in r mean more than for lower values.

THE SIGNIFICANCE TEST FOR r WHEN N IS SMALL

The rather high accuracy of the foregoing transformation technique provides a reliable method for testing whether a correlation coefficient calculated from a small sample signifies real correlation in the supply from which it is drawn. Warning has been given that, when ρ is zero and N is quite small, the sampling distribution of r although symmetrical is sufficiently platykurtic to make the use of a normal curve inadvisable. Transformation to z_r yields a curve more nearly normal in form which may well be used in preference.

In 1921, however, Fisher [6] provided another transformation in terms of which the appropriate probability may be secured without any approximation whatever. This transformation is as follows:

$$t = \frac{r\sqrt{N-2}}{\sqrt{1-r^2}}. \tag{5}$$

Detailed study of the symmetrical leptokurtic sampling distribution of t must be reserved for another occasion. Its probability integral has been rather fully tabled by Student,[7] who formulated the distribution for other purposes in 1908. Present interest in it lies solely in its provision of the precise sampling probability of r when ρ is zero and N is small. For this purpose, in lieu of giving Student's tables in full, Fig. 39 has been prepared. The lines of this chart trace the 10, 5, 2, and 1 per cent sampling probability values in r for values of N from 4 to 20. Locating the point (r, N) on this chart corresponding to any given sample value, one is able to determine whether the probability of the given value of r arising from an uncorrelated supply falls short of or exceeds 10, 5, 2, or 1 per cent. These percentages cover the critical zone in most tests of significance of a correlation coefficient.

[6] *Vide supra.*

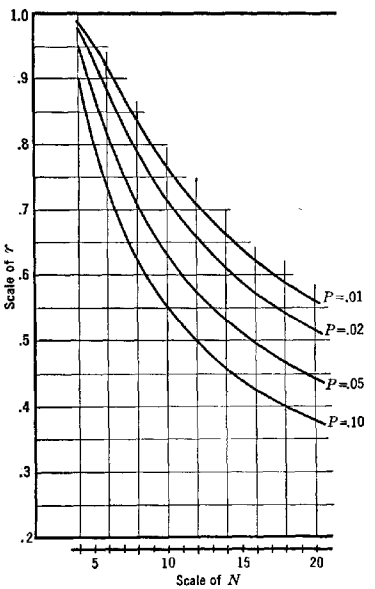
[7] Student. New tables for testing the significance of observations. Metron, 5 (No. 3): 18–21 and 26–32. 1925.

By way of illustration one may test the significance of the correlation coefficient +0.6125 based on 12 observations.

(1) Reference to Fig. 39 shows that the probability of this r being exceeded by random sampling errors alone is between 5 and 2 per cent. The precise value determined from "Student's" table is 0.0344.

FIGURE 39

PROBABILITY ZONES THAT THE DEVIATION OF r FROM ZERO MAY BE A RANDOM SAMPLING ERROR



(2) Using Fisher's logarithmic transformation, the relative deviate of z_r is $z_r\sqrt{N-3}$. Thus one finds $k_z = 2.1387$. From Appendix I, $P = 0.0325$.

(3) The coarser approximation of assuming the r distribution to be near enough to normal gives $k_r = r\sqrt{N-1} = 2.0314$, and $P = 0.0422$.

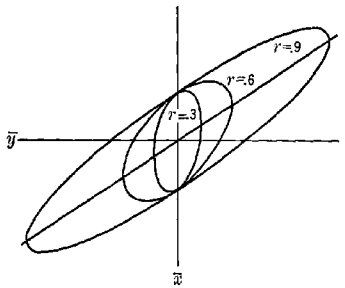
Thus the z_r technique gives very nearly the precise value. The $r\sqrt{N-1}$ technique yields somewhat too high a value of P, an error on the con-

servative side so far as claiming significance is concerned. As N increases, the discrepancy between the results given by the three methods decreases, becoming essentially negligible beyond N equal to 20.

In bringing to a close this brief discussion of sampling errors in the correlation coefficient, it seems pertinent to remind the reader that the correlation coefficient taken alone does not fully define any normal surface. Associations between the same variables having the same corre-

FIGURE 40

REPRESENTATION OF 3 NORMAL SURFACES HAVING IDENTICAL REGRESSION LINES AND RESIDUAL VARIATIONS IN PREDICTING y FROM x, BUT WIDELY DIFFERENT CORRELATION COEFFICIENTS



lation coefficient in samples of different origins may differ widely in their regression equations and residual variations. Likewise, surfaces may have precisely the same regression line and residual variation about that line in predicting one variable from the other, yet have very different correlation coefficients. Figure 40 is presented to illustrate the latter possibility. In the three surfaces represented therein by contours, the regressions of y on x are identical, and $\sigma_{y.x}$ is the same throughout. The values of r_{xy} are, however, 0.3, 0.6, and 0.9. The surfaces were readily constructed through selection of suitable values of σ_x and σ_y yielding the desired likenesses and contrasts.

The investigator depending on the correlation coefficient as a descriptive statistic for comparative work should not fail to secure thorough appreciation of such points as the above.

CHAPTER 12

PROPORTIONS AND PROBABILITY

Experience flowing from oft-repeated experiments has led to general establishment of the belief that, when an ordinary coin like those of the currencies of the English-speaking peoples is suitably tossed [1] and allowed to come to rest, the likelihood of one particular face of the coin being uppermost is not appreciably different from that of the other face being uppermost. Using such experience as a basis for forecasting the outcome of future experiments, one commonly states that the probability of the one event (e.g., head face uppermost) is equal to the probability of the other event (tail face uppermost). A third possible event, namely, that of the coin finally resting on its edge, may be recognized. However, any experiment yielding such a result may be rejected empirically as having no bearing with respect to the issue of the relative probabilities of "heads" and "tails."

This very familiar experiment is quite useful as a basis for the establishment (*a*) of certain definitions of wide scope, and (*b*) of consequences of a general nature flowing from those definitions. First of all there arises the concept of probability, the likelihood of an event occurring in the future based on experience of occurrence of that event and its alternatives in the past. Probability is belief founded on evidence. Probability may be given approximate numerical value by counting the events of finite experience and deriving the ratio of the count N_x of the specified event, x, to the count N_T of all the events wherein the specified one might have occurred. In the notation just given, N_T obviously equals the sum of N_x and the count N_y of the events y which are alternative to the event x; N_T is the total number of events wherein x might have occurred. The degree of approximation of this ratio to the true probability becomes closer, of course, as N_T increases. The true probability is precisely defined by infinite experience only.

If h is the event of a "head" in a coin-tossing experiment, t is the alternative event of a "tail," and we exclude any "coin-on-edge" event

[1] By this is implied a procedure of tossing or shaking which is free of particular motions or "tricks" calculated to bias the outcome of the experiment.

as belonging to an inappropriate experiment, then the "probability of heads," p_h, is given by the equation

$$p_h = \frac{N_h}{N_T} = \frac{N_h}{N_h + N_t}.$$

Again, if m designates male sex and f female sex, then the probabilities of male and female sex at birth as derived from a finite experience involving T births are:

$$p_m = \frac{N_m}{N_T}, \quad \text{and} \quad p_f = \frac{N_f}{N_T},$$

wherein $$N_T = N_m + N_f.$$

It is known from extensive experience that, under proper conditions, both p_h and p_m approximate to the numerical value of 0.5. The probability of securing the ace face (or any one specified face) in the random tossing of dice is commonly believed to be one-sixth, a value that would be expected on theoretical grounds for unbiased dice. The probability of germination of seed when placed in a proper environment is well known to depend on the kind and condition of the seed. Because of these factors it may vary over a wide range. Seed-testing laboratories are continually engaged in experiments to ascertain the percentage germination in submitted samples. Such determinations serve as estimates of the probability of germination appropriate to each supply of seed from which the sample is drawn.

These so-called probabilities are merely expressions of proportions in fractional form. A ratio of frequencies is involved in each case, the numerator of the ratio being by definition a part of the denominator. Thus the probability scale exists only between the limits of zero and unity. Also, each value of p on the scale is a pure number; it has no dimensions of measurement. The scale itself is a fundamental one, and may be designated for convenience as the standard scale of probability.

Proportions derived from frequencies are of very common occurrence in the analysis of biological data. They form the logical bases of comparative and interpretive work in many cases. Descriptions of the relative frequency of occurrence of events play a most important role in the field of vital statistics. Designated as *rates*, these numerical expressions will receive more extended consideration in the following chapter. At the present moment it is of interest to recognize merely that they form a class of proportions which may be accommodated to the standard scale of probability.

Although it is convenient in mathematical work to do so, it is not necessary that probabilities be expressed as fractions on the standard

probability scale. To many minds, simple whole numbers are more readily comprehended than fractions, and so probabilities are frequently quoted colloquially as percentages or as odds, while rates are commonly given as per 100, per 1,000, per 10,000, or per 100,000, according to the magnitudes in general of the type of proportion under consideration. Fundamentally, however, all are simple proportions derived from the corresponding magnitudes on the standard scale from zero to unity.

Reference so far has been confined to proportions derived from some known experience. It is by no means necessary to restrict our definitions in this manner. Probabilities may arise equally well from purely theoretical considerations, or even from arbitrary conjecture. While the latter may have but little value, theoretically determined proportions based upon sound reasoning embracing all known determinative factors may be intrinsically preferable to a value based on rather limited experience. In that which immediately follows, we shall use theoretically derived probabilities rather freely, on this account. The reasoning will be quite general, however, for all proportions, and will be equally applicable to all situations that are in accord with the fundamental hypotheses made.

Let p designate the proportion of favorable events in occurrences which may be either "favorable" or "adverse," and let q designate the proportion of the adverse events. The terms favorable and adverse as used here are not intended to imply quality, but serve as convenient designations of simple alternatives; adverse includes all events which are not to be designated as favorable. Then it follows irrefutably that $p + q = 1$ under such circumstances in all experiences. Let it be assumed now that there is available a large number of perfect "coins," each with its center of gravity lying precisely in that plane which bisects the rim. If the two faces are separately recognizable and designated as "head" and "tail" faces, then one can think of no logical reason why, in random tossing, one face should appear more frequently than the other. Designating one face as favorable, one may theoretically conclude that, for such coins, $p = q = 0.5$.[2]

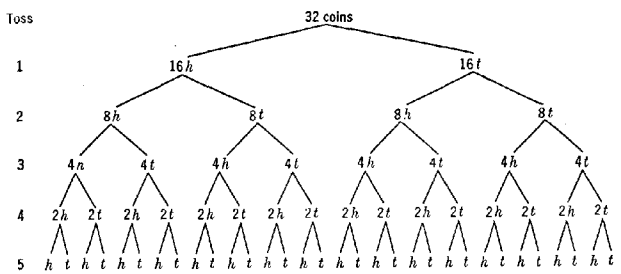
Consider now the actual random tossing of one such coin. In the first toss the result would either be one head or one tail. The coin cannot divide itself. Let the coin be tossed again. As the tossing is repeated under the ideal conditions of randomness, each toss will be uninfluenced by what preceded, and the proportions of heads and tails will approach statistically the limits

$$p = q = 0.5.$$

[2] The probability of heads in tossing an ordinary coin is likely to be an irrational fractional number slightly greater or less than 0.5, depending on the character of the coin.

Let us follow this assumption to its logical consequences when an infinitely large group of perfect coins are considered to be repeatedly thrown and classified after each throw. It will be convenient for the present illustration to write our expectancies in terms of a finite number of 32 coins. If these coins are tossed in a perfectly random manner, the coins then classified according to the uppermost face when they come to rest, then thrown again as subgroups, classified to the finer groups, and so on, the experiment may be anticipated in its results by the following ideal scheme:

SCHEME A



No real experiment with only 32 coins is ever likely to give such an ideal division between heads and tails at each toss. However, one wishes to represent here in terms of a limited number of coins the proportions that would be true if the number was infinitely great. After the fifth toss one would theoretically have 32 subdivisions of one coin each, one-half of them heads and one-half tails. The present major interest is to determine from the diagram the history of each coin in its successive falls. This is given in every case by the corresponding channel of the diagram. It may be seen, if the rather tedious analysis is undertaken, that no two histories are exactly alike, nor is any other type of history possible than is given by the diagram. Such sequences of events are known in mathematics as *permutations*. It follows from our ideal diagram that every possible permutation will occur with equal frequency at any given number of tosses. However, if the order in which the head and tail faces appear is ignored, then certain *combinations* will be observed to occur more frequently than others. If the superscripts to the letters h and t are used arbitrarily to indicate the number of heads or tails respectively shown in the history, then one may summarize all the combinations up to and including each successive toss by series of terms as follows:

SCHEME B

Toss	Combinations and Actual Frequencies Derived from Scheme A.
1	$16h^1 + 16t^1$
2	$8h^2 + 16h^1t^1 + 8t^2$
3	$4h^3 + 12h^2t^1 + 12h^1t^2 + 4t^3$
4	$2h^4 + 8h^3t^1 + 12h^2t^2 + 8h^1t^3 + 2t^4$
5	$1h^5 + 5h^4t^1 + 10h^3t^2 + 10h^2t^3 + 5h^1t^4 + 1t^5$

Each type of combination is represented above as h^xt^y, where x and y are respectively the number of heads and tails in the combination. The actual frequencies of occurrence of the different combinations are in each case the numerical coefficients of h^xt^y. Thus $10h^3t^2$ means that, on the average after the fifth toss, 10 coins out of 32 will show an experience of having fallen heads up 3 times and tails up twice.

Eliminating the superscripts and numerical coefficients of unity as being understood, and dividing the numerical coefficients of the several terms for any one toss by the common factor so as to reduce the frequencies to their simplest numerical proportions, one may write the respective combinations with their relative frequencies of occurrence as:

SCHEME C

Toss	Combinations and Their Relative Frequencies of Occurrence	Equivalent Binomial Expression
1	$h + t$	$(h + t)^1$
2	$h^2 + 2ht + t^2$	$(h + t)^2$
3	$h^3 + 3h^2t + 3ht^2 + t^3$	$(h + t)^3$
4	$h^4 + 4h^3t + 6h^2t^2 + 4ht^3 + t^4$	$(h + t)^4$
5	$h^5 + 5h^4t + 10h^3t^2 + 10h^2t^3 + 5ht^4 + t^5$	$(h + t)^5$

The reader will probably recall enough elementary algebra to recognize the similarity between these expressions for the combinations and the expansions of the binomial expression $(h + t)^n$, wherein n corresponds to the number of the toss, i.e., the number of events in each group. It may, in fact, be construed as the function of the binomial theorem to provide a systematic scheme of evaluating the relative number of each of the possible combinations of two independent events in multiple association.

In the foregoing reasoning a factor of the utmost importance has been the assumption of *independence* of the events in any combination. Coins may be tossed improperly, or dice so shaken that each result is not wholly independent of its associates in the combination. Professional gamblers are aware of and commonly use devices to interfere with the free play of chance. It is desirable for our purposes to have a rigid definition of independence, one readily amenable to mathematical test.

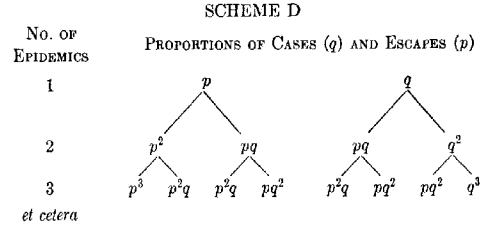
Let p_1 be the probability of an event and let p_2 be the probability of another event which may occur in combination with the former. Then the two events are independent of one another if the probability of joint occurrence of those events equals the product p_1p_2 of the probabilities of their separate occurrence. The truth of this statement, which comprises the usual mathematical definition of independence, may readily be made apparent in terms of an actual example. Let there be 3 men to every 2 women in a large office force, every member of which is given an apple. If 10 per cent of these apples are infested with grubs, and the apples are given out indiscriminately, then ideally 10 per cent of the men would receive infested apples and also 10 per cent of the women would be equally unfortunate. What is the probability that the combined event, a woman with a good apple, will occur? Since only 40 per cent are women and 90 per cent of the apples are good, 36 women on the average out of every 100 employees will receive good apples, if sex and goodness are independent; that is, the appropriate probability is 0.36, the product of the separate probabilities, 0.4 and 0.9.

Having thrown the concept of independence into this perfectly general form of a mathematical definition, one may proceed to apply it in a wide array of problems that may immediately suggest themselves. Returning to the coin-tossing experiment, if p is the probability of "heads," q that of "tails," then $p \times p \times p \times q \times q = p^3q^2$ is the probability of the permutation of 3 heads followed by 2 tails in tossing 1 coin 5 times. But we have seen that there are 10 different permutations, all equally likely to occur, giving this same combination of 3 heads and 2 tails. Is 10 then the appropriate factor by which to multiply p^3q^2 in order to calculate the probability of the *combination* of 3 heads and 2 tails?

Let us now proceed to apply the concepts of proportion and independence to a complete generalization of the reasoning underlying Scheme A. Expectancy being indicated in terms of proportion instead of frequency, and the number of events which may only be favorable or adverse being infinite, the combinations may be developed as before. This time p and q may have any fixed values subject to the limitation $p + q = 1$, and instead of tossing coins we may imagine repetitions of independent occurrences. Let us assume that a non-contagious, non-lethal, non-immunizing disease organism periodically sweeps the earth uniformly, and that in each epidemic the proportion of cases of disease developed is a constant q. Then p is the proportion of non-affected persons in each epidemic. In the first epidemic there will be a proportion q of cases and a proportion p of "escapes." In the second epidemic, since having had the disease before is stipulated to be without effect in preventing a second case, there will again be the proportions q and p of cases and escapes, but

the proportions p^2 and q^2 will respectively have escaped or developed the disease twice, while pq and qp proportions will have become cases once only. With a third and succeeding epidemic the same type of reasoning will hold, leading to the general plan as below.

SCHEME D



Since the number of the epidemic is the number of cases and escapes in each combination, we have the following generalized summary:

SCHEME E

No. of Events in Combination	Probabilities of Each Combination
1	$p + q$ $= (p + q)^1$
2	$p^2 + 2pq + q^2$ $= (p + q)^2$
3	$p^3 + 3p^2q + 3pq^2 + q^3$ $= (p + q)^3$
4	$p^4 + 4p^3q + 6p^2q^2 + 4pq^3 + q^4$ $= (p + q)^4$
.	. . . . .
.	. . . . .
.	. . . . .
n	$p^n + np^{n-1} + \cdots + npq^{n-1} + q^n$ $= (p + q)^n$

Thus the expansion of the binomial $(p + q)^n$ gives directly:

(1) the types of combinations of 2 alternative independent events in n experiences; and

(2) the numerical value of the probability of each combination when the values of p, q, and n are substituted in each term and the product secured.

Summarizing, one may state that, if p is the probability of a favorable event, q the probability of its alternatives collectively, and if these events are mutually independent in occurrence, then a group occurrence of n events of which r are favorable and $n - r$ adverse will have a probability of occurrence of kp^rq^{n-r}, where k is the appropriate coefficient defined by the binomial expansion. All one needs is a suitable rule for determining k in order to apply the reasoning to any practical problem.

The application of this proposition to the solution of problems in biology has proved most fruitful in testing independence of alternatives and in providing a mathematical graduation of actual data. One may be interested in such alternatives as survival or non-survival, sex of offspring, health and disease, success or failure, or a host of such problems wherein p and q may be hypothecated or determined from actual records. For instance, if sex in the human race is determined solely by chance factors, then the expansion of the binomial with appropriate values of p and q will give the proportions in which families of the various possible combinations of male and female children will occur in the general population. If p equals the probability of completed normal development of the human fetus, q equals that of the alternatives collectively, then the evaluation of $6p^5q$ will give the probability of 5 normal births out of 6 pregnancies if abnormalities are purely chance phenomena.

The expansion of the binomial may be written in general form as follows:

$$(p+q)^n = p^n + np^{n-1}q + \cdots + \frac{n!}{(n-r)!r!}p^{n-r}q^r + \cdots + npq^{n-1} + q^n, \quad (1)$$

where $n!$ (called factorial n) represents the product of all the integers from 1 to n. Numerical evaluation of the general term in this expression will give the probability of occurrence of $n - r$ favorable and r adverse events in a total group of n events. For those cases where one wishes to write down the full expansion of a binomial, it is convenient to remember the following characteristics of the series.

(1) Each term consists of the product of a numerical coefficient, a power of p, and a power of q.

(2) The first term is always $1p^nq^0 = p^n$.

(3) In succeeding terms the powers of p decrease by unity in regular order, while those of q increase in like manner, until the final term, $1p^0q^n = q^n$, is reached.

(4) The product of the numerical coefficient and the power of p in any term, divided by 1 *more* than the power of q, gives the numerical coefficient of the following term. Thus, the numerical coefficient of the second term will be n, that for the third term will be $\frac{n(n-1)}{2}$, and so forth.

Following these simple rules, one may write down immediately any expansion, for example,

$$(p+q)^5 = p^5 + 5p^4q^1 + 10p^3q^2 + 10p^2q^3 + 5p^1q^4 + q^5.$$

Remembering the general term of equation (1) will, however, enable one to evaluate any single term without having to write out all the preceding terms.

It may be noted that the number of terms in every binomial expansion exceeds the power involved by unity. The possible groupings of any size n must comprise types in which the favorable event takes values from *zero* to n, i.e., there must be $(n + 1)$ terms.

Probabilities of group occurrences derived from application of the binomial theorem obviously depend in part for their validity on the accurate knowledge of the value of p, the probability of the favorable event. Determination of the probability in favor of any given event may be a matter of theoretical reasoning in some cases, but it is frequently to be expected in biological investigations that its value must be based upon actual results. A value of p determined from a finite set of observations must be expected to be influenced by sampling errors. Another sample of the same size from the same general population would not necessarily yield exactly the same value of p. Such repeated samples would yield similar values of p only to a certain number of significant figures.

The dependability of p as defined by a finite experience will be analyzed when this discussion turns in a later chapter to the sampling errors of proportions. One may be content at this point to recognize that values of p determined by recorded experience rather than theory are trustworthy only to roughly about one-half the number of digits as those comprising the size of sample from which they are calculated. In a total of 5,017,632 human births considered, Geissler found 2,582,914 to be males. From these data one may calculate the proportion of the "favorable" events (males) as 0.514768. Since N here is a seven-figure number the derived proportion is, roughly speaking, stable under sampling errors to three or four places of decimals. It would seem foolish, therefore, to carry more than four figures in using this proportion as a probability of male sex at birth in computing binomial evaluations applicable to other sets of data.

Absolutely crucial to the correctness of any probability derived from application of the binomial theorem is the necessity that the combinations of favorable and adverse occurrences must be determined solely by chance. The appearance of each event must be absolutely independent of influence from those with which it is associated. Failure of observed results to accord (within sampling errors) with expectancy given by the binomial theorem must lead to the conclusion that either (*a*) the value of p used was incorrect, (*b*) the events comprising the groups did not arise independently of one another, or (*c*) both the above are true.

DISCRETE FREQUENCY DISTRIBUTIONS FROM THE BINOMIAL

One may pass directly to expected frequency of any definite group type in a finite experience involving N groups in all, by multiplying the group probability by N. This is merely reversing the procedure of securing proportions from observed frequencies. The expansion of the binomial $N(p + q)^n$ results in a series of $(n + 1)$ frequencies which correspond to the possible group combinations. The successive group types in the series of combinations are completely defined by the number of one element in the combination. This number proceeds from zero to n (or from n to zero) and defines the scale of a discrete variable. Therefore the expansion of $N(p + q)^n$ establishes a discrete frequency distribution which defines theoretical expectancy for the variable, provided, of course, that the value of p used is correct and the events are truly independent. Such distributions have considerable practical value; they also are of great theoretical interest. The following tabulation of calculations accompanying the solution of the problem indicated will provide a guide to the arrangement of work in a binomial expansion giving both group probabilities and group frequencies.

Seeds of Pima cotton were planted in 1,120 hills, each hill being sown with 6 seeds. After 2 weeks, 3,320 seedlings were found to form the surviving crop. If germination of the seed and survival of any seedling are independent of the fate of other individuals with which it is associated in the hill, what frequency distribution of seedlings per hill would theoretically be anticipated? Let p be the probability of survival. Then

$$p = \frac{3{,}220}{6{,}720} = 0.47917,$$

and

$$q = 0.52083.$$

Evaluation of the binomial $1{,}120\ (p + q)^6$, p and q having the magnitudes just given, will provide the desired distribution. This expansion is given in Table 21, wherein five decimal places are retained in all steps leading to the group probabilities. This follows a readily justifiable empirical rule that the number of decimal places carried should be 1 more than the number of digits in N. The theoretical frequency distribution (final column) will then be adequately accurate to whole numbers. It should be borne in mind that the distribution so secured is that having the p of the observed distribution, and it properly ignores in this phase of the work the dependability of p as a basis of estimation appropriate to other series.

It should be noted that the theoretical frequencies as calculated must be expected to embody fractional parts. They are related to one another strictly in the ratio of the group frequencies in an infinitely large experience. The fractional parts are often eliminated in practical work for convenience, but in so doing one is likely to obtain a total differing by one or two units from the value of N. This would occur in the present

TABLE 21

EXPANSION OF THE BINOMIAL FOR SEEDLINGS PER HILL

Number of seedlings per hill r	Theoretical probability formula kp^rq^{n-r}	p^r	q^{n-r}	Group probability kp^rq^{n-r}	Expected frequency Nkp^rq^{n-r}
6	$1p^6q^0$	$p^6 = 0.01210$	$q^0 = 1$	0.01210	13.6
5	$6p^5q^1$	$p^5 = 0.02526$	$q^1 = 0.52083$	0.07894	88.4
4	$15p^4q^2$	$p^4 = 0.05272$	$q^2 = 0.27126$	0.21451	240.3
3	$20p^3q^3$	$p^3 = 0.11002$	$q^3 = 0.14128$	0.31087	348.2
2	$15p^2q^4$	$p^2 = 0.22960$	$q^4 = 0.07358$	0.25341	283.8
1	$6p^1q^5$	$p^1 = 0.47917$	$q^5 = 0.03832$	0.11017	123.4
0	$1p^0q^6$	$p^0 = 1$	$q^6 = 0.01996$	0.01996	22.4
				0.99996	1,120.1

illustration if the theoretical frequencies were rounded to the nearest whole numbers. Perhaps the simplest solution to this problem is to adjust the largest frequencies in order by 1 each so as to obtain the correct total; that adjustment introduces the least effect of modification error on the results.

THE BINOMIAL SERIES AND THE NORMAL CURVE

The discrete frequency distribution arising from the expansion of any binomial $N(p + q)^n$ may be described in terms of its moments just as in the case of the continuous variables for which that procedure has previously been developed. The mean and standard deviation of the expected number of surviving seedlings per hill given by the binomial in the foregoing problem are calculated in Table 22. The values $\mu'_x = 2.878$ and $\sigma_x = 1.223$ are secured.[3] The frequency distribution may be observed in the table to be somewhat skew. One might proceed further to secure the third- and fourth-moment coefficients in like manner. However, all these computations may be circumvented in favor of very simple rela-

[3] Parametric notation is used because of the theoretical nature of the distribution.

tionships that exist between all the moment coefficients and the values of p, q, and n.

The binomial frequency distribution is given in algebraic form in Table 23, together with the first and second moments about zero for each group. Derivations A and B show that the mean numbers of favorable

TABLE 22

CALCULATIONS FOR MEAN AND STANDARD DEVIATION OF THE THEORETICAL EXPECTANCY FOR SEEDLINGS PER HILL

x	f	fx	fx^2
0	22	0	0
1	123	123	123
2	284	568	1,136
3	349	1,047	3,141
4	240	960	3,840
5	88	440	2,200
6	14	84	504
	1,120	3,222	10.944

$\Sigma x = 3{,}222.$ $\quad \mu_x' = 2.878.$

$\Sigma x^2 = 10{,}944.$ $\quad \dfrac{\Sigma x^2}{N} = 9.771.$

$\sigma_x^2 = 1.495.$ $\quad \sigma_x = 1.223.$

and adverse events are respectively np and nq, and that the standard deviation in both cases is $\sqrt{npq}$.

Derivation A: The mean number of successes is given by dividing the total of column (3) by N. Column (3) is composed of only n elements influencing the sum, the first member of the $n + 1$ entries being zero. Of these n elements, Nnp is a common factor. When this factor is withdrawn the remaining series is the binomial expansion of $(p + q)^{n-1}$. Therefore the total is $Nnp\,(p + q)^{n-1}$. But $(p + q)^{n-1}$ equals unity, as any power of $p + q$ does. Thus the mean of the series reduces to

$$\mu_x' = \frac{Nnp}{N} = np. \tag{2}$$

Derivation B: The standard deviation of number of successes per group is given by the usual equation,

$$\sigma_x = \sqrt{\frac{\Sigma x^2}{N} - \mu_x'^2}.$$

Now Σx^2 becomes the sum of the products of f and x^2 in the table notation. In arithmetic calculation this is secured by forming the column of fx^2 values. In the algebraic work of Table 23 we use another arrangement. Since

$$fx^2 = fx(x - 1) + fx,$$

and column (3) gives the fx products, column (4) is arranged to give the $fx(x - 1)$ products. Thus the sum of column (3) *plus* the sum of column (4) yields Σx^2 of the definitive equation for the standard deviation. The

TABLE 23

ALGEBRAIC DETERMINATION OF MOMENTS FOR THE MEAN AND STANDARD DEVIATION OF A BINOMIAL FREQUENCY EXPANSION

(1) Number of success x	(2) Frequency f	(3) fx	(4) $fx(x-1)$
0	$N q^n$	0	0
1	$N npq^{n-1}$	$Nnp\, q^{n-1}$	0
2	$N\frac{n(n-1)}{2} p^2q^{n-2}$	$Nnp\,(n-1)pq^{n-2}$	$Nn(n-1)p^2\, q^{n-2}$
3	$N\frac{n(n-1)(n-2)}{2(3)} p^3q^{n-3}$	$Nnp\frac{(n-1)(n-2)}{2} p^2q^{n-3}$	$Nn(n-1)p^2\,(n-2)pq^{n-3}$
.	.	.	.
.	.	.	.
.	.	.	.
n	$N\, p^n$	$Nnp\, p^{n-1}$	$Nn(n-1)p^2\, p^{n-2}$
Total	$N(p+q)^n$	$Nnp(p+q)^{n-1}$	$Nn(n-1)p^2(p+q)^{n-2}$

reader will see that a factor of $Nn(n - 1)p^2$ is common to all products in column (4). The sum of column (4) reduces to $Nn(n - 1)p^2\,(p + q)^{n-2}$, which is $Nn(n - 1)p^2$, since $(p + q)^{n-2}$ is again unity. Therefore

$$\Sigma x^2 = Nn(n - 1)p^2 + Np,$$

and

$$\frac{\Sigma x^2}{N} = n(n - 1)p^2 + np$$

$$= n^2p^2 - np^2 + np.$$

Since $\mu_x' = np,$

then $\sigma_x^2 = n^2p^2 - np^2 + np - n^2p^2$

$= np(1 - p)$

$= npq.$

That is, $\sigma_x = \sqrt{npq}.$ (3)

When p and q are reversed, the standard deviation will be the same for y' the number of failures.

The third- and fourth-moment coefficients may be derived in like fashion. The final equations are:

$$\mu_3 = npq(p - q), \tag{4}$$

and $$\mu_4 = npq + 3np^2q^2(n - 2), \tag{5}$$

from which it follows that

$$\beta_1 = \frac{(p - q)^2}{npq}, \quad \text{or} \quad \gamma_1 = \frac{p - q}{\sqrt{npq}}, \tag{6}$$

and $$\beta_2 - 3 = \frac{1 - 6pq}{npq} = \gamma_2. \tag{7}$$

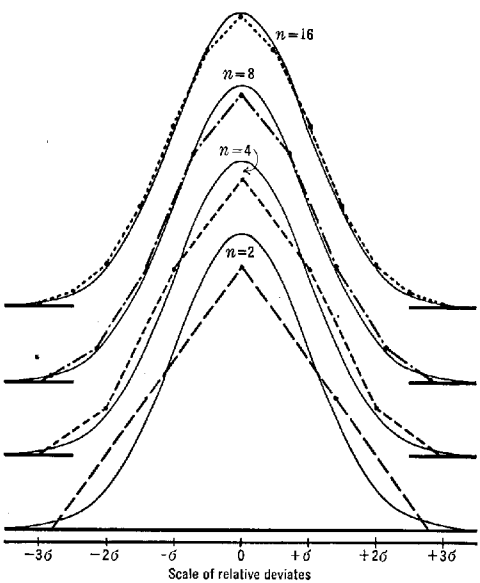
From these values it is plain that, when p is not equal to q, the binomial distribution is skew, the skewness increasing for any fixed n with increasing disparity between p and q. However, whatever the values of p and q, as n increases the skewness and kurtosis both decrease and approach the limiting values of zero, the normal curve values. This transition of the discrete binomial series toward the normal curve is a most intriguing one theoretically, and of very great importance to the simplification of many practical problems involving the binomial expansion. We shall therefore reconsider the transition immediately.

The theoretical range of the discrete binomial distribution is always from zero to n; that is, the range increases with n. For any fixed value of p, the mean and standard deviation also increase with n. If the distributions for increasing values of n are rescaled in terms of relative deviates, then the change in form of distribution with p fixed but n increasing may be studied geometrically. Under this scheme, as n increases, the ordinates defining frequency at the successive points will be drawn closer together until, in the limit of n approaching infinite magnitude, the distribution will become continuous. But, as $n \to \infty$, skewness and kurtosis of the distribution vanish, so the limiting distribution is the normal curve. The transition to the normal curve is fairly rapid when

$p = q = 0.5$. Indication of this is given in Fig. 41, wherein the binomial distribution is represented by frequency polygons for n equal successively to 2, 4, 8, and 16, all distributions being scaled in terms of relative deviates. The corresponding normal curve is drawn with each series. As the disparity between p and q increases, so the transition starts from increasingly skewed forms but moves nevertheless to the same normal curve as the limiting form. The skew binomial distributions with p equal to 0.95, 0.9, and 0.7, and the symmetrical distribution with $p = 0.5$, are given in Fig. 42, wherein n is held constant at 16 and the original scale from zero to 16 is preserved.

FIGURE 41

TRANSITION OF THE SYMMETRICAL BINOMIAL SERIES INTO THE NORMAL CURVE AS n INCREASES

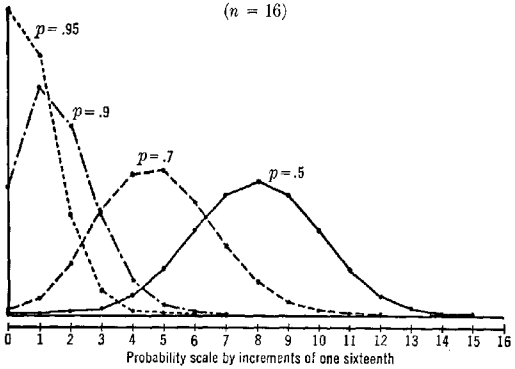


Binomials involving large powers are troublesome to expand, not only because of the great number of terms, but also because of the small probabilities which the single terms yield. Since the binomial distribution approaches the normal curve as n increases, and this curve is fully tabled, it is a great convenience to use the normal curve whenever it provides an

adequate approximation. No general statistical specification should be given for the adequacy of an approximation, since the importance of the error introduced should be judged on non-statistical grounds. However, the following empirical rule introduces only slight relative errors and may be considered as a first guide. If np (or nq if q is less than p) equals 20 or more, the normal curve may replace the binomial distribution for all but the most exacting practical problems. Lesser values of np, down to $np = 10$ quite reasonably, may indeed be accepted as usually providing adequate approximation. This rule is based upon the degree of approach of the binomial beta coefficients to the normal curve values.

FIGURE 42

TRANSITION OF THE BINOMIAL SERIES, DEFINED BY $(p + q)^n$, TO SYMMETRICAL FORM WHEN n IS FIXED AND p APPROACHES q ($n = 16$)



It must be borne in mind that frequency in a continuous distribution is given by area corresponding to a given range, whereas in the discrete distribution it is given by the ordinates alone at given points. If frequencies for binomial terms are to be approximated from corresponding areas in a normal curve, the range to be used in the normal curve must be widened from the binomial point or points by one-half unit on one side, or on both sides, as the problem demands.

THE POISSON SERIES

It has been noted that, when n is small and p is not equal to q, the binomial series forms a skew frequency distribution. Just as the normal curve provides a very simple basis of approximation to the binomial fre-

quency distribution under specified conditions, so the skew binomial which arises under certain other conditions is given satisfactorily by the exponential series,

$$Ne^{-m}\left(1 + m + \frac{m^2}{2!} + \frac{m^3}{3!} + \cdots + \frac{m^x}{x!}\right), \tag{8}$$

where e is the constant 2.71828, the base of natural logarithms. This series, first invented by Poisson in 1837, is generally designated as the Poisson exponential series. It may be shown mathematically that the Poisson exponential series increasingly approximates the binomial series as $p - q$ approaches unity and n is increased sufficiently that $m = np$ (or nq, whichever is smaller) remains finite but not necessarily large. Being much simpler to calculate, the Poisson series is often used in place of the binomial when $(p - q) \to 1$ and m is, say, between 0.1 and 10. For a given value of p, m varies directly with n, and as the latter is increased the Poisson series, together with the binomial series, approaches the normal curve as a limit.

Like the normal curve, the binomial series has two independent parameters if N is accepted as unity in both cases. They are n and p, for q is explicitly defined as $1 - p$. For the corresponding normal curve the parameters are $\mu' = np$ and $\sigma = \sqrt{npq}$. The Poisson series, on the other hand, has but one parameter, m, which is taken as the product np when the Poisson series is being used as an approximation to the binomial. The standard deviation of the Poisson series may be shown to be equal to $\sqrt{m}$; that of the binomial is $\sqrt{mq}$. Thus the binomial can transform precisely to the Poisson series only in the limit of $q \to 1$. They are, however, remarkably alike near this limit, and the Poisson series is so much easier to calculate that it is widely used as an adequate approximation.

It would be quite erroneous to infer from these remarks that the usefulness of the Poisson series is confined to its approximating character to the binomial series. Although this exponential series was first reached as a limiting form for the binomial, it is available for application to any observed discrete distribution starting at zero for which the mean and squared standard deviation are sufficiently in agreement. In this connection it is widely used as a graduating function for discrete distributions because of its very simple application. One may illustrate its use in this connection in terms of one of the experiments of Student in graduating the frequency distribution of cell counts with a haemacytometer.[4] Student used m equal to the observed $\bar{x}$ as his parameter in the Poisson series. One might choose alternatively to use m equal to s_x^2. Since in

[4] Student. On the error of counting with a haemacytometer. Biometrika, 5: 351–360. 1907.

the observed series $\bar{x}$ does not equal s_x^2, and discrepancy between these must be expected through errors of random sampling in finite series, the question arises which is the better one to use. Since the sampling errors of $\bar{x}$ are considerably less than those of s_x^2, the reason for choosing the former as a definition of m is obvious.

In the Poisson series as defined in equation (8), note that the first

TABLE 24

ACTUAL AND FITTED POISSON SERIES FOR THE DISTRIBUTION OF NUMBER OF ERYTHROCYTES ON HAEMACYTOMETER SQUARES

Data from Student [4]

Number of cells per square x	Observed frequency	$\frac{m}{x}$	Poisson series	
			Theoretical frequency	Rounded frequency *
0	103		106.6	107
1	143	1.3225	141.0	140
2	98	0.66125	93.2	93
3	42	0.44083	41.1	41
4	8	0.330625	13.7	14
5	4	0.2645	3.6	4
6	2	0.220416	0.8	1
	400		400	400

$\Sigma x = 529.$ $\bar{x} = 1.3225.$

$\Sigma x^2 = 1{,}213.$ $\frac{\Sigma x^2}{N} = 3.0325.$

$s_x^2 = 1.2835.$ $s_x = 1.1329.$

m for the Poisson expansion taken as equal to $\bar{x}$.
Then Ne = antilog of $\{\log N - m \log e\} = 106.6$.

term is Ne^{-m}. Each successive term may be derived from its predecessor, as in Table 24, by multiplying it by $\frac{m}{x}$, x being the scale value of the variable. These terms total to N, of course. To secure probabilities instead of frequencies, the first term may be evaluated as e^{-m}, from which point the procedure is the same. Logarithms are usually necessary to calculate e^{-m}, since m rarely proves to be a whole number.

* Since taking the nearest whole number gives a total frequency of 401, we have dropped 1 from the highest frequency. Making the first frequency 106 might be preferred by some.

CHAPTER 13

THE PROPORTIONS OF VITAL STATISTICS

One of the many specialized domains of application of the statistical method, the field of vital statistics deals with the organization and analysis of data pertaining principally to the health, length of life, and reproduction of more or less large units of population. Vital statistics then assumes an important role in the broad movement aiming to describe the well-being and social organization of mankind. Many variables of diverse nature enter into consideration in this field of inquiry. Some of them, such as age and physical dimensions of man, are susceptible of fairly precise determination through measurement. The majority of the variables of interest, however, are subject only to qualitative description. In the introductory Chapter it was suggested that the complexities involved in describing the health of human beings must result in attainment only of a rather crude division into a few general classes. In contrast to this, sex is determinable with accuracy in all but an inappreciable proportion of individuals. With all such qualitatively defined variables the description is made by categorical designation, the accuracy of the classifications changing from those of high degree to others of but vague implication. . In general, the individuals to be described are recognized to fall somewhere in a wide variation scheme, in which two or more subclasses are definable as parts of the range which may be assembled to form the total range of variation. Thus an individual may be recognized as being alive or dead with respect to the state of vitality; or being in good, medium, or poor health; as belonging to one of many well-defined nationalities or several intermingling races; as being male or female with respect to sex; *et cetera*. Descriptions of individuals in such terms involves the recognition of qualities innate to them at the time of observation. Such qualities may well be designated as attributes of the individuals. The statistical analysis of such data pertaining to a population is often called the statistics of attributes, to distinguish the type from the statistics of measured variables.

The qualities of the population with which the student of vital statistics is concerned are in the main related to the general subjects of mortality, morbidity, and natality. The appearance or existence of the

quality may be designated as the *event* or *state*. For purposes of terminology, one may use the term *event* in a general sense here to include also those occurrences more properly designated as states. Thus one may refer to records of death, sickness, and births as records of events, vital statistics being concerned with description of the frequency of occurrence of such events in a comparative way.

There is not suitable opportunity to concern oneself herein with the interesting problems of securing the basic data comprising the counts of the events and of the populations in which they may take place. Preliminary reading on this subject may be found in textbooks concerned principally with vital statistics, among which may be cited that of Pearl entitled "Medical Biometry and Statistics." Suffice it to say that the main burden of accumulating such records falls naturally to the lot of civil governments, most of which provide special administrative agencies to care for this work. The difficulties that face their officials in securing full and accurate records are formidable, and in some degree insuperable. The element of inaccuracy in the basic data must therefore never be lost sight of. Sometimes the data are very crude. On occasion they are so selectively incomplete or inaccurate that it is questionable whether they are better than no information at all, if not worse. In many other situations they are highly dependable, and of course all gradations may be expected between these extremes.

The statistical methods of resolution of data, on the other hand, may be, and indeed usually are, highly refined. One may pertinently choose to give here a simile previously suggested. Like a scalpel in the hands of a surgeon, the statistical method depends for the practical effectiveness of its work on the skill of the person applying it, not only in his knowing how to use the instrument but also in understanding the material to which it is applied. The student of vital statistics must be doubly careful to scrutinize his basic data for the accuracy of description and coverage involved before applying to them the formal procedures of statistical analysis. It is primarily with that method of analysis that one is concerned herein.

RATES

The vital statisician, and many another engaged in social studies, is assigned the task of describing the frequency of occurrence of specified events in a population. The description required must be in terms capable of ready comparison. The task might well be designated as one of describing the *force of incidence* of the event. By way of example one may consider the intensity with which death, regardless of the cause, impinges on a population. What was the force of mortality in the state of Maine in 1930, for instance, and how does it compare with that in the

state of Montana in the same year? To say that the number of deaths in the two areas were respectively 11,082 and 5,440 is inadequate description, for obviously if the force of mortality was the same the deaths would be reasonably proportional to the respective populations, other things being equal. This very simple idea of proportionality provides immediately a clarification of what is specifically implied by the term "force of mortality;" it is described by the proportion that deaths form in a population in a given time period. Such proportions, measuring the force of incidence of an event, may be calculated for different types of events wherever the basic data are available. Expressed as numbers on scales ranging from zero to some suitable power of 10, these proportions are commonly known as *rates* in vital statistics.

Two numerical elements are necessary for the determination of true proportions convertible to rates. They are: (*a*) a count of the event of reference, and (*b*) a count of the population in which the event can take place. The reader may be asked to note carefully this specification of a particular logical relationship between these two counts. A tendency to destroy the descriptive value of the ratio will be introduced if the population used is other than the one exposed to risk of occurrence of the event. Since death may come to anyone, the whole population is the proper base for establishing the general death rate from all causes. In measuring the force of mortality through pregnancy, however, only those in the pregnant state are exposed to the risk of occurrence of the specified event; to use any broader population must introduce an element of confusion into the description.

RATES AS APPROXIMATING EXPRESSIONS

The ideal rate describing incidence of an event would result from the "follow-up" of a population through a period of exposure to the event. Thus, ideally, an annual death rate from all causes would give the proportion that deaths formed among a population living at a point of time and enumerated with respect to the incidence of death throughout the following year. Let l_x be the number of individuals forming the living populations at a point of time x. Let l_{x+1} be the number of survivors from that population 1 year later. Then the number of deaths d_x in the year of observation would be given by

$$d_x = l_x - l_{x+1},$$

and the proportion of deaths or "per capita death rate" for the year would be definable as

$$q_x = \frac{d_x}{l_x}.$$

Such a rate as q_x, logically used as a measure of expectancy applicable to a comparable population, would measure the probability of death facing that population in the ensuing year.

The securing of records of individual histories such as are called for above is obviously out of the question as a civil administrative procedure in large populations. The clerical work involved in assembling annual reports by all living persons would alone forbid the practice.

Bearing in mind the practicability of procedures aiming to permit description of the vitality characteristics of a population, it might be demanded merely that specified events such as births, deaths, and incidence of readily communicable diseases be reported to civil administrations. In this connection it may be noted that streams of migration continuously permeate civil jurisdictional boundaries, more or less steadily changing the population within them. Also, those who die cease to be reporters of events to any terrestrial government; it must be demanded of others to report on them and on the sick. Thus only those births, cases of sickness, and deaths occurring within specified areas may at best be expected to be reported. These may take place in part among infants and immigrants as new arrivals during any time interval, similar events among emigrants from the area being recorded in some other jurisdiction.

The factors mentioned above must be considered in the construction of an appropriate mortality rate. The population is changed more or less progressively by operation of those factors, and the deaths are properly referable only to that changing population. An estimate is therefore needed of the average population in the interval. Assuming the changes to have taken place fairly uniformly throughout the period, the logical population to accept for construction of the mortality rate becomes that at the mid-point of the time interval, that is, as of July 1 if the period is the calendar year. This population must usually be estimated from interpolation in, or extrapolation of, the population growth curve as indicated by the periodic censuses and accessory bases of inference. The general death rates of vital statistics are in practice so computed, and other rates fall in the same class. They are logical approximations to the ideal figure which is defined by purely theoretical considerations.

The death rate used above, describing the force of mortality regardless of cause as it impinges on a general population, has been discussed primarily as a specific illustration of an expression approximating to an ideal proportion. Another mortality rate commonly used in medical statistics may be contrasted with it from this point of view. In the general population, individuals are likely to come under continued

observation because of some disease that has been contracted by them, or because of surgery or other treatment to which they have submitted. Such persons may be classified according to the ailment from which they suffer, and be followed until recovery or death takes place. The proportion that deaths from the condition form among the cases is a mortality rate determined from recorded histories and therefore accords with the foregoing specification of an ideal rate. Commonly known as a *case fatality rate*, it is not an approximating expression in the sense of the general death rate previously considered. The case fatality rate may be completely accurate for the type of case coming under observation provided it may be accepted that death was the consequence of the ailment prescribed.

Some conditions which may terminate in death of the patient, and for which a rate description of the force of incidence of death is desired, do not come under observation in a way permitting of the construction of an accurate case fatality rate. An outstanding example is that of deaths resulting from the puerperal state. Although cases of pregnancy leading to death of the mother are recorded, the total number of pregnancies to which the deaths should properly be related in construction of the ideal rate is not readily determinable with accuracy. In lieu of this most appropriate denominator for the expression it is customary to use the number of live births through the same period of time. One need not discuss the degree of approximation involved here beyond remarking that the base is accepted as being better than the alternative ones which may be generally available in official records. The derived rate is known as the *maternal mortality rate*. Its deficiencies as an approximation are compensated for to some degree by fairly dependable comparability of the rate from one region to another.

The three foregoing types of mortality rates must serve to indicate that the rates of vital statistics may vary considerably with respect to attainment of the ideal form of description given by the proportion that a part forms of the whole. Whereas some rates approach closely to desired perfection in description, others may approximate the ideal expression of proportion rather coarsely. Practical considerations not uncommonly forbid the attainment of ideal goals in science. Rates may often be constructed from components which do not bear the mathematical relationship to one another which would be desired by precision objectives. They may nevertheless be logical approximations having the quality of comparability in the same way as the unattainable ideal ratios of a part to the whole.

THE PERTINENCE OF RATES TO PLACE AND TIME

The counts of populations and events incident to them, from which descriptive rate expressions may be formulated, pertain to national or political subdivisions of the earth's area. Specification of the area of reference for any rate is, of course, a necessary part of the description. Likewise, definition of the calendar period in the history of the population through which the event is counted is essential to allocation of the rate, for most rates have been changing progressively through time. Each numerically determined rate must pertain to some area of reference and period of historical time. The italicized words in the following illustrative statement are essential to interpretation of the numerical rates given: The case fatality rate from scarlet fever *in the city of Providence* has fallen from 15.1 per cent in *the quinquennium 1884-1888* to 2.7 per cent in *1918-1922*.[1]

TEMPORAL RATES

Although all rates pertain to some time period, historically speaking, two distinct classes of rates exist with respect to the functional influence of length of time involved on the rate itself. This may conveniently be illustrated in terms of two of the mortality rates previously discussed, the "general death rate" and the "case fatality rate." The former is the ratio of the count of an event *through a time interval*, to the count of the appropriate population *at a point of time*. Time does not cancel out in the ratio; the numerical value of the rate will itself be roughly proportional to the length of the interval through which the event is counted. Such rates should always be adjectively qualified by an expression of the length of time interval involved. Thus death rates may be annual, quarterly, monthly, *et cetera*, and their numerical values under constant conditions will be proportional to the length of the interval. The annual death rate in Minnesota in 1930 was 10.00 per 1,000, whereas that for the two months of June and July combined was 1.62 per 1,000, which is approximately one-sixth of the annual rate. The population used in both periods is identical (that of July 1), whereas the number of deaths in each period is very different.

Case fatality rates may be taken as representative of the class of ratios in which both numerator and denominator represent counts through the same time interval. Under such conditions time cancels out to leave the rate as a pure number. The quinquennial case fatality rates for scarlet fever just given for the city of Providence may also be defined as the average annual rates over the same periods. A five-year

[1] *Vide infra*, Table 25.

span was used merely to strike a more representative value for the general period. The basic data are given in Table 25, wherein the case fatality rates by single years may be compared with those of the five-year periods.

TABLE 25

REPORTED CASES AND DEATHS FROM SCARLET FEVER IN TWO 5-YEAR PERIODS FOR THE CITY OF PROVIDENCE, RHODE ISLAND

(Data from Chapin *)

Year	Cases	Deaths	Case fatality rate per 100
1884	538	57	10.6
1885	383	38	9.9
1886	237	30	12.7
1887	848	153	18.0
1888	361	79	21.9
1884–8	2,367	357	15.1
1918	367	18	4.9
1919	528	11	2.1
1920	533	14	2.6
1921	319	6	1.9
1922	161	2	1.2
1918–22	1,908	51	2.7

THE NUMERICAL POPULATION BASE

Another problem peculiar to the numerical specification of a rate is that of the standard size of population referred to. One may conveniently return to the question of the death rates for 1930 in Maine and Montana. The basic data are given in Table 26.

TABLE 26

CALCULATION OF GENERAL DEATH RATES, MAINE AND MONTANA, 1930

(1) Area	(2) Deaths in 1930	(3) Population † July 1, 1930	(4) Proportion of deaths	(5) Death rate per 1,000
Maine.....	11,082	798,140	0.0139	13.9
Montana..	5,440	537,330	0.0101	10.1

* Annual Reports of the Superintendent of Health of the City of Providence for the years 1916–1922. Providence: The Oxford Press. 1923.

† Estimate made by assuming that census figures of January 1, 1920, and April 1, 1930, are dependable and that the average three-month population change between the dates continued to July 1, 1930.

Verbal statement of the proportions given in column (4) might lead one to say that, for every person in Maine, 0.0139 of a death occurred in 1930. The proportion given is wholly analogous to an average, and as such its fractional character is not inappropriate. However, a fraction of a death lacks ready comprehensibility to the lay mind. This difficulty is partly circumvented by changing the proportion from a "per capita" basis to a "per 1,000" basis, or to any other suitable basis yielding whole numbers of deaths. It is in such form that the ratios are designated as *rates*. One or two decimal places are frequently retained after the whole numbers even in the rate form when the greater numerical precision provided by more significant figures is considered desirable. The "per 1,000" basis is accepted by convention as a standard for the annual rate of deaths from all causes because it yields small whole numbers easy of comprehension. On the same grounds, decennial death rates if used might well be given as per 100. Other classes of rates require different standard bases to achieve these same ends, as may be noted later in this chapter. In the absence of an acceptable standard basis common to all rates, it is obviously absolutely essential that the basis be defined whenever rates are given in numerical form.

SPECIFIC RATES

Why, one may well ask, did the force of mortality in 1930 appear to be so much higher in Maine than in Montana? Was it much healthier to live in Montana? Acknowledging that this does not seem reasonable, one might turn attention to the examination of factors influencing these rates, factors that so far have been overlooked in our preliminary consideration. Were the populations of the two states comparable with respect to their composition so far as that may affect the force of mortality? The risk of imminent death is well known to be higher for the old than for the young. Perhaps there was a greater proportion of older people in Maine. The general risk of death may perhaps differ also by broad racial groups (white versus negro, for instance), by occupation or urbanization, by sex, and so forth. Many such questions arise for consideration in seeking a solution of discrepancies observed in the comparison of two or more areas or time periods.

It is of value to notice that all these suggestions emanate from recognition of the existence of relatively more homogeneous elements into which the general population may be divided. Death rates, for instance, may be computed within such specific classes of the general population. Two or more areas may be compared in terms of these specific class rates with greater refinement than is possible when heterogeneity is

ignored. Such rates for relatively homogeneous subgroups in the population may be designated for purposes of distinction as *group specific rates.* By comparison the rate for the heterogeneous population as a whole is called a *crude rate,* which indeed it is.

An age specific rate is one calculated for a specified age group in the population. There will, of course, be as many of these rates arising for one general population as there are age groups recognized in that population. To consider every year of age separately would give a large number of classes, larger than is warranted by the change in the risk of death. It is because of this that broader classification than by single years of age is often used. The coarser grouping chosen is usually an irregular one, for experience shows that the risk of death does not change uniformly over the life span. This may be seen in the table that follows.

Within each of such primary groups a further subdivision into racial classes, say white and colored, might be made. If there were 10 age groups, then there would be 20 age-color classes, 40 age-color-sex classes, and so forth. Now an age specific rate is relatively crude compared to an age-color specific rate, and so on. Thus a progression in degrees of specificity has been established by the concept of refinement in homogeneity of recognizable subclasses of any general population. The original crude rate corresponds to what might be called a "zeroth order" of specificity, single-factor specific rates being of the first order of specificity, two-factor specific rates being of the second order, and so on. According to this terminology, the order of specificity of a rate is the number of independent criteria of classification which are used to specify the ultimate classes into which the population is to be divided.

Discussion of this subject in terms of actual data must be confined herein merely to illustration of a first-order specification. In considering the contrast in the Maine and Montana crude death rates above, it is reasonably obvious that the most likely source of differentiation which would help account for the disparity in the crude death rates would be the age composition of the populations. One may proceed to investigate this by calculating the age specific death rates. The basic data and derived rates are given in Table 27, where a conventional age classification is used. The population estimates by age have been derived by assuming that the relative distribution by age at the census date, April 1, persisted without change to July 1.

Comparison of the rates of death from all causes within the specified age classes in Table 27 reveals a striking similarity for the two states considered. This is in sharp contrast with the disparity between the two crude rates given against "all ages" at the foot of the table. Obviously, the difference in the crude rates is not due to a greater force of

mortality in Maine, age for age, but to the difference in age composition of the two populations. Recognition of specific age classes within the populations has thus contributed greatly to clarification of the basis of an initially surprising discrepancy. It has, however, introduced a new problem of comprehension in that the single pair of crude rates is now replaced by 16 pairs of age specific rates. The large number of differences so arising is not readily assimilable as a whole beyond recognition

TABLE 27

AGE SPECIFIC AND TOTAL DEATH RATES IN MAINE AND MONTANA, 1930

Age	Maine			Montana		
	Estimated population	Deaths in 1930	Death rate per 1,000	Estimated population	Deaths in 1930	Death rate per 1,000
→1	14,442	1,227	84.96	9,834	583	59.28
1	14,594	130	8.91	9,522	81	8.51
2	15,573	80	5.14	9,832	43	4.37
3	15,186	49	3.23	9,878	43	4.35
4	15,365	57	3.71	10,202	32	3.14
5– 9	79,858	148	1.85	53,998	90	1.67
10–14	74,183	104	1.40	56,402	108	1.91
15–19	68,796	153	2.22	50,141	146	2.91
20–24	60,674	224	3.69	43,758	166	3.79
25–29	53,197	183	3.44	38,199	171	4.48
30–34	52,700	230	4.36	35,480	152	4.28
35–44	101,358	552	5.45	82,696	534	6.46
45–54	90,495	980	10.83	63,476	702	11.06
55–64	72,597	1,476	20.33	37,209	783	21.04
65–74	46,690	2,433	52.11	20,236	959	47.39
75→	22,432	3,056	136.23	6,467	847	130.97
All ages	798,140	11,082	13.88	537,330	5,440	10.12

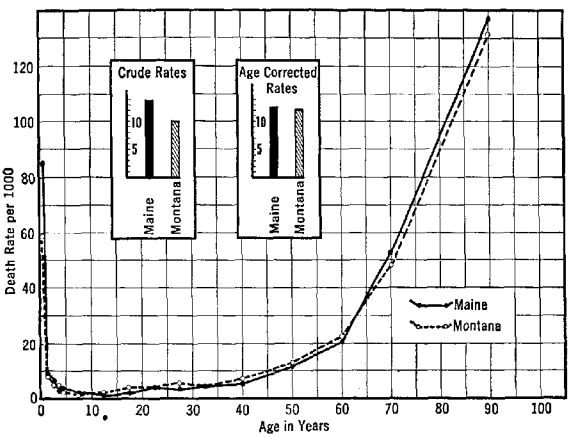
of the fact that, in general, they are relatively small. Graphic comparison of the age specific rates by means of two "trend lines" in Fig. 43 materially aids comprehension of this latter point. However, the attendant problem of *greater sampling instability* of the specific rates, because they are necessarily based on smaller numbers of people, arises for consideration. The significance of the difference between any pair of specific rates may be expected to be of a quite different order from that of the difference between the two crude rates.

CORRECTED RATES

There will, in general, be as many specific rate comparisons as there are subclasses designated. In the foregoing illustration the broad age groupings in the regions where the rates are not changing markedly[2] are desirable because they give greater sampling stability to the rates secured. Where the total populations are not large, or the number of

FIGURE 43

CRUDE, AGE SPECIFIC, AND AGE CORRECTED DEATH RATES. MORTALITY FROM ALL CAUSES IN MAINE AND MONTANA, 1930



specific classes is considerable, the populations per specific class may be so reduced as to render some or all of the specific rates of questionable value, whereas the crude rates may have been fairly dependable from the point of view of sampling errors. In view of such possibilities, if not from other bases of reasoning, one may naturally raise the question: What would the total rates have been if the populations had not differed in composition with respect to the character made specific?

Despite apparent reasonableness of this question on first consideration, careful analysis readily shows that it is unanswerable without the use of populations other than the true ones in which the deaths took place. Answers to the question must tend to carry one away from

[2] The very broad group, 75 →, is imposed by lack of more detailed published data on age at death in this range.

reality, for hypothetical populations must be invoked. The acceptability of the transfer will, of course, depend on the usefulness of the information which the answer yields.

The basic inquiry in constructing corrected rates is to ascertain the total number of deaths that would have resulted if the group specific rates for each observed population had prevailed on some other population used as a standard. Such standard populations may be structurally theoretical but derived from actual population counts of past censuses, or the actual populations themselves may be used. Also, those populations may, on the one hand, be quite independent of, or on the other hand derived from, the ones of reference in the specific rate correction problem undertaken. The principle of keeping as close to reality in selection of the standard population as conditions will permit has much to commend it. To return to the data on deaths in Maine and Montana, this principle may be observed in either of two ways:

(*a*) The age specific rates of one state may be applied to the population of the other to determine an "expected number of deaths" for comparison with those actually occurring. This procedure obviously will be possible in two directions if there are but two states, or in $n(n-1)$ directions if there are n states, yielding much confusion if n is large.

(*b*) The age specific rates of each state may be applied to the combined population of both states to secure "expected deaths." With n states there will be only n such sets of "expected deaths."

When the general principle is considered, rather than the anticipated comparison of only two areas, method (*b*) above would be selected as the more satisfactory. It is applied to the comparison of Maine and Montana in Table 28, wherein the last two columns, calculated from the preceding three columns, give the "expected number of deaths" if the force of mortality by age characteristic of each state had afflicted the population of the two states combined. The total numbers of "expected deaths" are found to be 16,690 and 16,104 respectively, derived from the age specific rates of Maine and Montana. This rather close agreement in the pair of comparable figures stresses the very similar force of mortality that characterized the two areas when the age composition of their populations is accounted for. The number of deaths may in each case be converted to a rate by dividing by the total population of 1,335,470 in which they are calculated to occur. These total death rates, determined as 12.50 and 12.06 per 1,000 respectively, are known as *age corrected* (or *age adjusted*) *rates*.

It is unfortunately all too easy to overlook the hypothetical nature of corrected rates. To state from the calculations that 12.50 per 1,000 is the age corrected general death rate for Maine in 1930 is quite erroneous. The population used in arriving at the corrected rate was that of Maine and Montana combined. Other populations might have been

TABLE 28

CALCULATION OF AGE CORRECTED TOTAL DEATH RATES FOR THE COMPARISON OF MAINE AND MONTANA, 1930

Age group	Population			Observed deaths rates per 1,000		Expected deaths * under rates of	
	Maine	Montana	Total	Maine	Montana	Maine	Montana
→1	14,442	9,834	24,276	84.96	59.28	2,062	1,439
1	14,594	9,522	24,116	8.91	8.51	215	205
2	15,573	9,832	25,405	5.14	4.37	131	111
3	15,186	9,878	25,064	3.23	4.35	81	109
4	15,365	10,202	25,567	3.71	3.14	95	80
5– 9	79,858	53,998	133,856	1.85	1.67	248	224
10–14	74,183	56,402	130,585	1.40	1.91	183	249
15–19	68,796	50,141	118,937	2.22	2.91	264	346
20–24	60,674	43,758	104,432	3.69	3.79	385	396
25–29	53,197	38,199	91,396	3.44	4.48	314	409
30–34	52,700	35,480	88,180	4.36	4.28	384	377
35–44	101,358	82,696	184,054	5.45	6.46	1,003	1,189
45–54	90,495	63,476	153,971	10.83	11.06	1,668	1,703
55–64	72,597	37,209	109,806	20.33	21.04	2,232	2,310
65–74	46,690	20,236	66,926	52.11	47.39	3,488	3,172
75→	22,432	6,467	28,899	136.23	130.97	3,937	3,785
Total	798,140	537,330	1,335,470	13.88	10.12	16,690	16.104
Age corrected total death rate per 1,000						12.50	12.06

chosen, yielding perhaps substantially different figures. This is illustrated in Table 29, wherein the theoretical life table population of 1910 for the original registration states is used as a standard. The age corrected general death rates therein derived from the Maine and Montana age specific rates are 14.98 and 14.43 per 1,000—figures substantially higher than those secured from Table 28.

* In the total population.

The numerical value of any corrected rate is not descriptive of any occurrence. The meaningful value is the relative magnitude of that rate when compared to some other corrected rate calculated on the *same* population. The ratios of the age corrected death rates for Maine and Montana derived from the two standard populations are given in Table

TABLE 29

CALCULATION OF AGE CORRECTED TOTAL DEATH RATES FOR THE COMPARISON OF MAINE AND MONTANA, 1930

Age	Observed death rates per 1,000		Life table population *	Expected deaths † under age specific rates of	
	Maine	Montana		Maine	Montana
→1	84.96	59.28	17,841	1,516	1,058
1	8.91	8.51	16,916	151	144
2	5.14	4.37	16,612	85	73
3	3.23	4.35	16,448	53	72
4	3.71	3.14	16,338	61	51
5–9	1.85	1.67	80,682	149	135
10–14	1.40	1.91	79,628	111	152
15–19	2.22	2.91	78,513	174	228
20–24	3.69	3.79	76,802	283	291
25–29	3.44	4.48	74,717	257	335
30–34	4.36	4.28	72,342	315	310
35–44	5.45	6.46	135,917	741	878
45–54	10.83	11.06	121,068	1,311	1,339
55–64	20.33	21.04	98,831	2,009	2,079
65–74	52.11	47.39	65,377	3,407	3,098
75→	136.23	130.97	31,968	4,355	4,187
Total	13.88	10.12	1,000,000	14,978	14,430
Age corrected total death rate per 1,000				14.98	14.43

30. These ratios are essentially identical despite the discrepancies in the age corrected rates provided by the two populations. The important feature revealed by the new rates is that the force of mortality for Maine, when adjusted for age, is only 4 per cent greater than that appropriate to Montana, whereas the crude death rate for Maine was nearly 40 per cent greater than for Montana.

* U. S. Original Registration States, 1910 Life Table.
† In the life table population.

RATES SPECIFIC FOR SUBDIVISIONS OF THE EVENT

The division of populations into more homogeneous subclasses often provides a crucial refinement in formulating rate comparisons between areas or points in time. Perhaps there is no more striking illustration of this than flows from the widely varying age specific death rates, ranging from less than 2 per 1,000 about 10 years of age to a theoretical limiting value of 1,000 per 1,000 at the end of the human span of life. In introducing rate expressions, death from any cause is a convenient event to choose because existence of the state of death is rarely open to controversy and it is an event to which all are exposed. This particular event and its companion morbidity are, however, conventionally subject to subclassification according to some so-called "cause" of the death or sickness. We must forego here any considera-

TABLE 30

COMPARISON OF AGE CORRECTED TOTAL DEATH RATES, MAINE AND MONTANA, 1930

Ratio	Standard population	
	Combined actual	Life table *
$\frac{\text{Maine}}{\text{Montana}}$	$\frac{12.50}{12.06} = 1.04$	$\frac{14.98}{14.43} = 1.04$

tion of the arbitrary nature of such subclassifications with their frequently great errors of judgment, and concern ourselves solely with recognition of a new category of specific rates, namely, those termed *cause specific rates.*

There is no difficulty in accepting cause specific rates as a descriptive type. A problem more or less peculiar to them, however, is to give them suitable numerical expression. This arises from the widely varying proportions which they take on a "per capita" basis. Some causes of sickness or death are common; others are extremely rare; and no one standard size of population is eminently suitable as a rate basis for all. In only very few cases, for example heart disease, or cancer and tumors collectively, is a population unit of 1,000 a suitable basis yielding small whole numbers for such rates, the point being that although large numbers are exposed relatively few are smitten. Accordingly, convention

* U. S. Original Registration States, 1910 Life Table.

for some time stipulated 100,000 as the base for numerical expression of cause specific rates.

Increasing control over certain diseases is already making the "per 100,000" too small a base where before it was adequate. A quarter of a century ago the annual death rate from typhoid fever in Minnesota was approximately 11 per 100,000, but for the last ten years it has been less than 1 per 100,000. Thus some cause specific death rates need to be quoted as "per million" or a higher power of 10 to preserve whole numbers in the rates. One must always be alert to ascertain the numerical population base on which cause specific rates are quoted. One should also preserve a wholesome regard for the effects which differences of nomenclature or classification procedure, frank or illusive errors of judgment, and changing pressure of medical opinions and public reaction may have on the comparability of cause specific mortality or morbidity rates between different areas or points of time.

SOME WIDELY USED RATES DEFINED

In closing these general comments on the proportions of vital statistics it is perhaps pertinent to remark that the diverse expressions of rate have been developed over a span of years without any organized plan of nomenclature or critical review of the different types which have come into existence. Probably the most important matters for the statistician to recognize in dealing with these descriptions are those listed below.

(1) They all correspond to simple proportions and, in statistical analysis procedure, fall directly in that class of measures.

(2) Their magnitudes are dependent on multiplying a numerical magnitude on the standard probability scale by some power of 10, that power being more or less conventionally set up for the particular type of rate in question. Without a clear statement of this multiplier, which is simply a scale transformation factor, rate expressions are useless.

(3) Many rates (but not all) are a function of time in the sense that the magnitude of the rate is proportional to the *length* of the time interval. Such rates to be meaningful must have an accompanying statement of the time interval involved.

(4) Most rates of vital statistics are expressions approximating to some ideal ratio which is not itself readily determinable. Understanding of the nature of this approximation is essential to thorough appreciation of the descriptive quality and usefulness of the rate involved, when comparative work is undertaken.

Some of the more commonly used rates are listed here for reference.

General death rate: the number of deaths irrespective of cause in a time interval, per 1,000 population at the mid-point of the interval.

Cause specific death rate: the number of deaths ascribed to the specific cause in a time interval, per 100,000 population at the mid-point of the interval.

Case fatality rate: the number of deaths from a specific disease, per 100 reported cases of that disease, both counts being made through the same specified interval.

Maternal mortality rate: the number of deaths directly attributable to pregnancy, per 1,000 live births, both counts being made in the same specified interval.

Infant mortality rate: the number of deaths in the first year of life, per 1,000 live births, both counts being made in the same specified time interval.

The *neonatal mortality rate* similarly considers deaths within the first month of life.

Birth rate: the number of live births in a specified interval, per 1,000 population at the mid-point of the interval.

This very crude rate is often refined to include in the denominator only women who are in the reproductive years of life (15 to 45, or 10 to 60).

Subdivision may be made for legitimate and illegitimate births, with appropriate denominators.

Stillbirth rate: the number of stillbirths, per 100 live births, both counts being made through the same specified time interval.

Morbidity rate: the number of cases of disease, in a given time interval, per 1,000 or per 100,000 population at the mid-point of the interval.

This rate is usually cause specific.

CHAPTER 14

SAMPLING ERRORS OF PROPORTIONS

It is a generalization of fairly wide acceptance that 10 per cent of all cases of typhoid fever may be expected under present conditions to terminate in death of the patient. This anticipation is based on considerable experience with reported cases of the disease, the case fatality rate derived from "follow-up" studies in large areas appearing in the last decade or two to have remained essentially unchanged. With a view to studying the sampling distribution of proportions in terms of familiar situations and actual numbers one may for present purposes accept this proportion of case fatalities as being correct.

Consider now 1,000 sets of 10 cases each of typhoid fever, chosen at random from a supposedly infinitely large supply. One may proceed on the basis of the binomial theorem to determine how often the number of recoveries per group of 10 cases must be expected to be 10, 9, 8, $\cdots$, zero. Let the probability π of recovery be accepted as 0.9; then the probability of death, $1 - \pi$, becomes 0.1. Expansion of the binomial

$$1{,}000[\pi + (1 - \pi)]^{10}$$

will give the expected numbers of groups with 10, 9, 8, *et celera*, recoveries per group. The arithmetic expansion is given in Table 31, wherein it may be noted that roughly one-third of the experiences with sets of 10 cases will show recoveries without any deaths, one-third with one death, and one-third with 2 or more deaths. These results are inevitable consequences flowing from the assumption that each set of 10 cases is assembled by random selection from a general supply having a case fatality rate of 10 per cent.

One may direct attention now to these sets of 10 cases each. The number of recoveries per set, taken as a proportion of the number in the set, defines the probability of recovery so far as each set *alone* is capable of defining that probability. These probabilities p, based on finite experience, are given in the final column of Table 31.

The reader will now discern the reason for using π as the probability of recovery in the original expansion. It is the assumed true or parametric value for the probability of recovery, the values p given by the samples of 10 being corresponding statistics derived from finite experi-

ence. The p values will be observed to differ from one another solely through errors of random sampling, for by definition the individuals comprising the samples were drawn at random. Thus the errors of random sampling in observed proportions, p, are explicitly defined by the binomial expansion based on the true proportion, π, of infinitely large experience. For samples of size 10, one may expect once or twice in

TABLE 31

DETERMINATION OF THE EXPECTED NUMBERS OF RECOVERIES IN SETS OF 10 CASES OF TYPHOID FEVER

Probability of recovery, $\pi = 0.9$

Recoveries per 10 cases	Group probability	Frequency per 1,000 sets of 10 cases	Recovery indicated by the set, in per cent
10	0.348678	349	100
9	0.387420	388*	90
8	0.193710	194	80
7	0.057396	57	70
6	0.011160	11	60
5	0.001488	1	50
4	0.000138		
3	0.000009		
2	0.000001		
1	0.000000		
0	0.000000		
Totals	1.000000	1,000	

1,000 sets of 10 cases that half the patients will die when the true probability of death is only 10 per cent. For the same reasons slightly more than one-third of all samples will show no deaths at all.

It has already been proved[1] that the mean *number* of "favorable events" in a binomial distribution is np, the standard deviation being $\sqrt{npq}$. Changing p to π to distinguish the parameter from the equivalent sample statistic, and dividing by the number of events per group (n) to reach the probability scale from the frequency scale, it follows that the equations

$$\mu_p' = \pi, \tag{1}$$

and

$$\sigma_p = \sqrt{\frac{\pi(1-\pi)}{n}}, \tag{2}$$

* Point of least relative error of modification to give total of 1,000.

[1] *Vide* page 176 *et seq.*

define the mean and standard deviation of p based on random samples of size n. When the size of sample is small, the individual terms of this discrete distribution may be evaluated without difficulty. When n is large, it has been shown in Chapter 12, the normal curve defined by equations (1) and (2) above will give adequate approximation to the distribution probabilities. The problem of sampling errors in probabilities has therefore been rather completely analyzed in these pages, the needed theoretical distributions being established.

THE STANDARD ERRORS OF PROPORTIONS AND RATES

The true standard deviation of the sampling errors in proportions p is defined in equation (2) above as being dependent on the value of π in the supply. Without knowledge of this value it is impossible to define that standard deviation precisely. One is therefore obliged to estimate π when it is desired to attach to the value of p yielded by a unique sample some measure of its sampling stability. Equation (1) above is of interest in this connection in that it shows that p is an unbiased statistic. *On the average*, p is equal to π. The best estimate of π given by a unique sample is therefore the sample value of p. Substituting this in equation (2), one secures an estimate of σ_p, which estimate must appropriately bear designation as the *standard error* of p:

$$\text{S.E.}_p = \sqrt{\frac{p(1-p)}{n}} = \sqrt{\frac{pq}{n}}. \tag{3}$$

Rates have been defined as true proportions or logical estimates of proportions expressed on a scale from zero to some specified power of 10. One might well designate the upper limit of that scale as the numerical population base for the rate. Let B be that base and R the rate. Then the basic proportion of the standard probability scale leading to the rate is

$$p = \frac{R}{B},$$

the rate being equal to Bp. It follows directly then that the standard error of any rate is defined by the equation

$$\left.\begin{aligned} \text{S.E.}_R &= \sqrt{\frac{R(B-R)}{n}} \\ &= B\sqrt{\frac{p(1-p)}{n}}. \end{aligned}\right\} \tag{4}$$

Approximate confidence intervals with respect to π may readily be

established on the basis of equation (3) provided that n is large enough to warrant use of the normal curve for the sampling distribution of p. The situation then becomes identical with that wherein a confidence interval is to be defined for the mean of a supply, starting from a sample mean and its standard error. The 95 per cent confidence interval for π will cover the approximate range from $(p - 1.96\,\text{S.E.}_p)$ to $(p + 1.96\,\text{S.E.}_p)$. Other confidence intervals may be established in analogous manner.

A practical illustration of the sampling trustworthiness of a proportion may be given in terms of the very large series of records of sex at birth accumulated by Geissler.[2] He found that among 5,017,632 human births the number of male offspring was 2,468,305. If p is the derived probability of male sex at birth, then

$$p = 0.514768,$$

and

$$q = 0.485232.$$

From this one determines that

$$\text{S.E.}_p = \sqrt{\frac{0.514768 \times 0.485232}{5{,}017{,}632}}$$

$$= 0.000223.$$

The 95 per cent confidence interval for locating the probability of male sex at birth that would be given by an infinitely large experience is therefore from 0.5143 to 0.5152. It would thus seem inadvisable to use the given p to more than three or four places of decimals in discussions of the probability in general of male sex at birth.

CONCORDANCE OF p WITH π

It has been noted that on the average p is identical with the parameter π. In individual samples, however, p must be expected to vary about π in a binomial distribution with standard deviation given by equation (2). One may therefore readily test whether a value of p given by a sample is consistent with any assumed value for π. If n is small, the binomial expansion may be used to find the probability that the deviation $p - \pi$ would arise solely through errors of random sampling. When n is large the normal curve probabilities may be used as providing a good approximation.

One may turn again to a specific problem for illustrative development of the principle. Diphtheria deaths reported for the state of Minnesota

[2] Arthur Geissler. Beiträge zur Frage des Geschlechtsverhältnisses der Geborenen. Zeitschr. K. Sächs. Statistischen Bureaus, 35: 1–24. 1889.

in 1935 show the following distribution by sex: male, 5; female, 11. It being assumed that the cases of diphtheria were equally divided between the sexes, what is the probability that a distribution of deaths between the sexes as uneven as this is due to chance?

The small number of cases involved in this problem makes direct expansion of the binomial quite feasible. In the absence of differential mortality rates between the sexes, the expected proportion of male deaths, among all deaths, would be defined by $\pi = 0.5$. Expansion of the binomial $[\pi + (1 - \pi)]^{16}$ will therefore yield the probabilities of each possible numerical combination of male and female deaths through random sampling. It is necessary now to examine carefully the question propounded, so that the appropriate probability may be derived from

TABLE 32

PROBABILITIES OF 5 OR FEWER MALE DEATHS IN GROUPS OF 16 PERSONS, WHERE $\pi = 0.5$

Number of deaths		Group probability
Males	Females	
0	16	1×0.5^{16}
1	15	16×0.5^{16}
2	14	120×0.5^{16}
3	13	560×0.5^{16}
4	12	$1{,}820 \times 0.5^{16}$
5	11	$4{,}368 \times 0.5^{16}$
	Total	$6{,}885 \times 0.5^{16}$

this binomial. Fewer than 5 male deaths out of 16 represents a greater departure from equality than is given by 5 alone, and therefore all combinations from zero to 5 male deaths show at least as great a departure as that specified. Also, the combinations involving 5 or fewer female deaths represent a departure from equality as great as that specified. The question fundamentally centers about the deviation from equality, not merely the ratio of 5 males to 11 females. Thus the sum of the probabilities given by the combinations involving 5 or fewer male or female deaths is required. Since the binomial is symmetrical, only the first six terms need be evaluated; double their sum will give the requisite probability.

Evaluation of the probabilities corresponding to the specified six terms of the binomial with $\pi = 0.5$ is given in Table 32. Since π is equal

to $1 - \pi$, then $\pi^r(1 - \pi)^{n-r}$ is equal in all terms to 0.5^{16}. The numerical coefficients of the successive terms may be written in directly, following the simple rule given earlier.[3] The total probability derived from the specified terms then becomes

$$P = 2[6{,}885\ (0.5)^{16}] = 0.21.$$

Thus slightly more than 20 per cent of the time, with samples of 16, a distribution of deaths between the sexes as uneven as that observed would arise solely through random sampling errors when the expected division is equal.

Another problem of like nature may be considered. A report on an outbreak of typhoid fever in Mankato, Minnesota,[4] in 1908, records 35 deaths from the disease among 511 cases. What is the probability that a departure as great as this from the general expectation of a 10 per cent mortality rate is due to chance?

The answer to this problem would be given precisely by expanding the binomial $[\pi + (1 - \pi)]^n$, where $\pi = 0.1$ and $n = 511$, the terms being summed from zero deaths up to and including 35 deaths. Such computation would be onerous indeed. The product of n and π in this problem is far in excess of 20, a value suggested in Chapter 12 as indicating that very close agreement exists between the normal curve and the true binomial distribution, higher values giving even closer agreement. Normal curve areas may therefore be used in this problem with confidence as providing very close approximations to the true binomial frequencies.

The continuous nature of the approximating curve contrasts here with the discrete character of the binomial distribution it is to replace. This difference should properly be adjusted for by recognizing that the total probability corresponding to 35 and fewer deaths in the binomial is equivalent to the relative area under the curve up to 35.5 deaths. Under the hypothesis being used, namely, that π is 0.1, 51.1 deaths would be expected. Therefore the normal curve, scaled in terms of number of deaths, is centered at 51.1 and has a standard deviation $\sqrt{n\pi(1 - \pi)}$ equal to 6.7816. The relative area below 35.5 deaths may be found in Appendix I after calculation of the relative deviate,

$$k = \frac{35.5 - 51.1}{6.7816} = 2.30.$$

One finds $P = 0.0214$.

[3] *Vide* page 172, also page 250.

[4] Journal of Infectious Diseases, 9: 410–474. 1911.

Only once in 47 experiences with sets of 511 cases each would deviation from expectancy as great as that observed be anticipated through sampling errors alone. Possibly in this epidemic the typhoid bacillus was somewhat attenuated, or not all the reported cases were of typhoid fever but some were milder ailments of somewhat similar symptoms. In any event, the validity of the hypothesis that the general mortality rate for the ailments dealt with was 10 per cent is open to grave suspicion.

Precisely the same conclusions must follow if the numerical expressions take the form of proportions instead of frequencies of deaths; 35.5 deaths out of 511 is a proportion (p) of 0.0695, the mean and standard deviation of the sampling curve now being defined by equations (1) and (2) above as

$$\mu_p' = 0.1,$$

and

$$\sigma_p = 0.01326.$$

Then

$$k = \frac{p - \mu_p'}{\sigma_p} = \frac{0.0305}{0.01326} = 2.30,$$

the value previously secured. The same value would follow if case fatality rates in percentage are used, the changes throughout being linear transformations of scale and having no influence on the pure number relative deviate finally secured for entering the table of the normal curve.

THE SIGNIFICANCE OF THE DIFFERENCE BETWEEN TWO PROPORTIONS

The foregoing problems deal with the deviation of an observed proportion from some specified value which is prescribed by theoretical considerations or based on a sufficiently wide experience to be accepted tentatively as the true proportion. A second class of closely related problems deals with the consistency of two proportions derived independently from finite samples, the parameters being unknown. Attention will now be given to such problems in terms of actual data.

The distribution between the sexes of both cases and deaths in the Mankato typhoid fever epidemic referred to above is given in Table 33. What is the probability that the difference in the case fatality rates for the two sexes is due to chance?

Note that this question does not deal with the deviations of the respective rates from any specified value; it is concerned solely with the difference between them. One is required to test the null hypothesis, and any assumptions made must be in accord with that hypothesis. Now the general sampling errors in rates appropriate to each size of sample

being dealt with are a function of π, the true or parametric proportion. An estimate of π is therefore demanded, and if the null hypothesis is true then the best estimate available will be given by pooling the experience for both sexes. In the "total" array in Table 33 the case fatality rate is given as 6.85 per 100, making 0.0685 the logical estimate of π to be

TABLE 33

CASES AND DEATHS FROM TYPHOID FEVER, BY SEX
MANKATO EPIDEMIC, 1908

Sex	Cases	Deaths	Case fatality rate per 100 *
Male	221	16	7.24
Female	290	19	6.55
Total	511	35	6.85

employed. The standard errors of sampling in the observed rates (R) will then become respectively:

n	$S.E._R^2$	$S.E._R$
221 (men).....	2.8872	1.6992
290 (women).....	2.2003	1.4833

The values of $n\pi$ do not fall far below 20, and normal curves may be accepted as good enough first approximations to the actual sampling distributions of the rates involved. The observed rates for men and women, of 7.24 and 6.55 respectively, may tentatively be presumed to belong in normal curves of mean 6.85 and standard errors as given above.

In establishing the sampling distribution of differences between means paired at random,[5] the solution of all random sampling difference problems originating in normally distributed statistics was established. It was there shown that, under the null hypothesis,

$$\mu_d' = 0,$$

and

$$\sigma_d = \sqrt{\sigma_x^2 + \sigma_y^2},$$

* Two decimal places are carried here solely for numerical accuracy in the derived computations.

[5] *Vide* Chapter 10, page 138 *et seq.*

where d is the difference between the two statistics $\bar{x}$ and $\bar{y}$. By transferring to the rates (R_m and R_w) and replacing true standard deviations by standard errors, it follows that

$$\begin{aligned} \text{S.E.}_d &= \sqrt{\text{S.E.}_{R_m}^2 + \text{S.E.}_{R_w}^2} \\ &= \sqrt{2.8872 + 2.2003} \\ &= 2.26. \end{aligned}$$

Now $$d = 7.24 - 6.55 = 0.69,$$

and the relative deviate of this difference in its normal sampling distribution is given by

$$k\,(=)\,\frac{0.69}{2.26} = 0.30.$$

From the tables in Appendix I,

$$P = 0.76,$$

indicating an entirely insignificant difference between the case fatality rates for the two sexes. That such a conclusion would be in order was

TABLE 34

TEST OF THE NULL HYPOTHESIS APPLIED TO AGE CORRECTED TOTAL DEATH RATES, MAINE AND MONTANA, 1930

Data derived from Table 27

	Maine	Montana	Total
Population	798,140	537,330	1,335,470
Expected deaths *	9,975	6,479	16,454
Death rate, R, per 1,000	12.50	12.06	12.32
S.E._R^2 †	0.0152	0.0226	
Difference in rates, d	0.44		
S.E._d	0.194		
k	2.27		
P	0.0232		

obvious from the fact that the standard error of each rate was considerably greater than the actual difference between the rates.

* Reduced to actual population bases.
† Based on $r = 0.01232$.

A comparatively small difference between rates based on large numbers of cases may, of course, be quite significant. By way of illustration one may take the age corrected total death rates derived from the Maine and Montana experiences in 1930. Table 34 gives the data and calculations pertinent to testing the null hypothesis in that investigation. It should be noted carefully that the population bases used in testing the significance of the difference between corrected rates must logically be no larger than those from which the basic age specific rates are derived. Therefore in Table 34 the "expected deaths" are reduced to the numbers appropriate to the recorded populations of Maine and Montana.

The probability of the difference between the age corrected total death rates in Table 34 being due to chance alone proves to be very small. One might well consider this significant and perhaps continue the study by correcting for sex, "urbanization," *et cetera*, as well as age, to determine whether the finer group specific forces of mortality are not more nearly identical. Since the principles of procedure have been set forth above, the analysis will not be continued herein. They may be undertaken by the interested reader in order to test his grasp of that which has been presented.

The foregoing study of sampling errors in proportions provides widely useful techniques for testing the concordance of proportions considered by pairs. It has been shown that the analyses may proceed in terms of proportions, rates, or actual frequencies, it being immaterial which form is adopted. Attention will now be given to the broader problem of the concordance of an array of pairs of proportions. The analysis will proceed for convenience entirely in terms of frequency.

CHAPTER 15

THE MEASUREMENT OF FREQUENCY DISCORDANCE

One of the great intellectual contributions of Gregor Mendel to biological science was that he foresaw and demonstrated the possibility of stating the results of biological experimentation in simple mathematical expressions. Realizing that forces of chance incidence obscured agreement between the results of his plant hybridization experiments and those he theoretically anticipated, he labored diligently at repetition of his work to minimize the effect of chance through having large numbers. Finally he was convinced and felt that others would be convinced of the adequacy of his theory on the basis of his great and precious store of evidence. Yet Mendel did not ultimately secure perfect agreement between the observed frequencies of occurrence of his phenotypes and those which he theoretically expected. The actual differences were relatively small. He attributed them to what the statistician would call "errors of sampling" and rested his hypothesis on the ground that the observational facts approached the theoretical conception more and more closely as the number of observations was increased.

Mendel lacked an objective criterion of the goodness of agreement between observed and theoretically expected frequencies. It was not until 1900, 34 years after Mendel published his work and in the year when it was rediscovered, that Karl Pearson provided a mathematical criterion to meet Mendel's need. Pearson was not concerned specifically with Mendel's problem, for he undoubtedly did not know of it when his memoir was written. He was considering such problems in general, and the discussion immediately following will be given with the same general outlook, illustrated with specific examples.

Considerable attention has already been directed to the fact that, if the probability π of an event is known, and if such events and their alternatives may occur independently in groups of size n, then the probabilities of the possible group combinations are given by the numerical values of the separate terms in the expansion of the binomial $[\pi+(1-\pi)]^n$. Such an expansion, however, yields ideal values to which actual observations can approximate only within the errors of random sampling. If two "perfect" coins be randomly tossed 4 times, the results will not necessarily be 2 heads, head and tail, and 2 tails, in the ratio of 1 : 2 : 1. Each toss will

yield one of these combinations, but the results of the second, third, and fourth tosses will be quite independent of what happened previously. In the long run, however, as the number of repetitions of the experiment is increased, closer and closer approximation to the theoretical ratio of 1 : 2 : 1 will be reached. Lack of exact agreement between the observed and theoretical frequencies must be expected purely through chance, and such discrepancies will not in any sense invalidate the theoretical frequencies.

Admitting that exact agreement between observational and theoretical frequencies should not be expected in any finite experience, one is faced with the difficulty of distinguishing between agreement which is satisfactory and that which is unsatisfactory. Any one person's judgment will probably not agree with that of all others. The theoretical expression may indeed fail to portray the true distribution law. This brings one face to face with a problem of the utmost importance. One must have an *objective* criterion of the closeness of agreement of the observed with the theoretical frequencies. Without such a test the critical worker may well remain hopelessly confused as to whether the observations fully substantiate the theoretical distribution of frequency or not.

Consider the general case of two series of frequencies distributed into identical classes, the one series arising from observation and the other from theoretical deduction. Within each class let o designate the observed frequency and c the calculated frequency, the total frequency for each series being N. Then $o - c$ will measure the absolute discrepancy in each frequency class. Let n' designate the number of such classes. Then the simplest possible summary of the discrepancies will be secured by adding the n' values of $o - c$. But

$$\sum^{n'}(o - c) = \sum^{n'} o - \sum^{n'} c$$

$$= N - N.$$

This summation obviously defeats the purpose; on the whole there will always be no discrepancy by this method.

The difficulty again is a matter of signs. It may be surmounted as before by considering the second moments instead of the first. The total discrepancy may be measured as $\sum^{n'}(o - c)^2$, a positive quantity of magnitude E, say. Now E will vary in the aggregate with the magnitude of the individual differences, and hence E would appear to fulfill the requirement of measuring the discrepancy between observation and theory. However, it suffers one severe limitation. The *importance* of any given

absolute discrepancy depends on the number expected. If theory should call for a frequency of 500 in a certain group and 502 cases are observed, the concordance is relatively very much greater than if 5 cases were expected and 7 were observed. Since the discrepancies in the n' classes must finally be summed, the measures of discrepancy should be in comparable terms. The frequency discordance should then be measured on a relative basis, that is, as $\frac{(o-c)^2}{c}$ or $\frac{(o-c)^2}{o}$.

In choosing between these two forms of measuring the relative discrepancy in each class, two considerations are of predominant importance. First, when c is small in magnitude o may conceivably often be zero solely through errors of sampling. In such cases the relative discrepancy would be indeterminate if $\frac{(o-c)^2}{o}$ is used. To any suggestion that c may be zero and o a greater magnitude, thus making $\frac{(o-c)^2}{c}$ indeterminate, the answer must be clear. If theory does not provide for a frequency that does occur, then the theory is obviously inadequate and it is foolish to attempt to measure concordance. Quite apart from this issue of determinability of the measure itself, there is the second consideration that the calculated frequency c represents a wider experience or broader conception than the frequency o shown by the finite sample; c should then form the logical reference value.

The value $\frac{(o-c)^2}{c}$ is then proposed as a measure, comparable from one situation to another, of the importance of discrepancies between theoretical and observed frequencies. One may with benefit consider this measure of frequency discordance in relation to a numerical example. If 6 coins were tossed at random 64 times, the number of "heads" being recorded for each toss, a frequency distribution could be prepared for the observed numbers of heads per toss. Now 6 coins, each with a probability of 0.5 of showing the "head" face after random tossing, must show an average distribution of heads in 64 tosses as given in the c column of Table 35. The column o in that table gives frequencies empirically selected as quite typical of the results of an actual tossing experiment. The differences $o - c$ vary from zero to 4, the derived elements $\frac{(o-c)^2}{c}$ varying from zero to a maximum of 1. The magnitudes of the latter elements in relation to the importance of the deviations $o - c$ comprise the matter of immediate interest. It will be profitable to study this relationship with care.

TABLE 35

ILLUSTRATING THE COMPARABILITY OF THE ELEMENTS $\frac{(o-c)^2}{c}$ AS MEASURES OF FREQUENCY DISCORDANCE

Heads per toss	Frequencies		Discrepancies $o-c$	$\frac{(o-c)^2}{c}$
	Calculated c	Observed o		
6	1	0	−1	1.0
5	6	8	+2	0.67
4	15	12	−3	0.6
3	20	24	+4	0.8
2	15	14	−1	0.07
1	6	5	−1	0.17
0	1	1	0	0.0
Totals	64	64	0	3.31

The following two points may well be noted from the data of Table 35.

(1) The three unit deviations (for 6, 2, and 1 "head") yield $\frac{(o-c)^2}{c}$ values inversely proportional to the expected frequency. That is, as the expected frequency goes up, the importance of any *fixed* deviation from it of observed frequency goes down. This was deliberately chosen as a requirement of the measure of frequency discordance.

(2) When the deviation $o - c$ is *proportional* to the expected frequency, as in the "4 heads" and "3 heads" groups, the values of $\frac{(o-c)^2}{c}$ are directly proportional to the expected frequencies. This most important quality of the measure of frequency discordance arises through squaring the deviation but retaining the denominator c in the first power. Just as a deviation of 2 from an expected 500 is of far less consequence than a deviation of 2 from 5, so a deviation of 200 from 500 is of far more consequence than a deviation of 2 from 5.

The measure suggested by Pearson therefore fulfills very simply and neatly the two outstanding logical requirements for a widely comparable measure of discordance between an observed frequency and its theoretical equivalent.

Establishment of an elemental measure of the lack of agreement between the frequencies pertinent to each class permits consideration of the ultimate problem, that of measuring the discordance in all classes as a whole. One wishes to know the probability that the deviations of observed frequencies from the theoretical in the distribution as a whole are attributable to random sampling effects alone. The elements $\frac{(o-c)^2}{c}$, each being proportional to the general importance of the deviations, may well be totaled for the entire set of classes. That sum forms the criterion designated by Pearson as χ^2 (called "chi squared"):

$$\chi^2 = \sum^{n'}\left[\frac{(o-c)^2}{c}\right]. \tag{1}$$

In any practical problem of measuring total frequency discordance, such as that presented in Table 35, for instance, χ^2 is a numerical quantity without dimension of any sort. It is a pure number derived from pure number frequencies. The scale of χ^2 is obviously limited at zero as far as smallness is concerned, but may assume very large values when there is great discrepancy between the observed and theoretical frequencies. It is the smallness, numerically speaking, of χ^2 which reflects the degree of concordance of observed and calculated frequencies. There is some limit above which values of χ^2 will indicate lack of satisfactory agreement between observation and theory. This limit, however, will not be any fixed number applicable to all situations, for χ^2 is a sum of basic elements and its magnitude must inevitably be a function of the number of those elements.

One may choose to calculate a quantity such as χ^2 with the sole objective in view of determining the probability that the total discrepancy it measures might be due to errors of random sampling. It is necessary then to ascertain the random sampling distribution of χ^2 derived from n' deviations so that one may know how large its value may become solely through the chance errors of sampling. This rather complex problem was solved by Karl Pearson as part of his classical memoir[1] establishing the criterion. Tables of the probability integral of χ^2 for values of n' from 3 to 30 were then prepared by Elderton,[2] from which it is possible to

[1] Karl Pearson. On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling. Phil. Mag., 50: 157–175. 1900.

[2] W. Palin Elderton. Tables for testing goodness of fit. Biometrika, 1: 155–163. 1902. See also Table XII in Part I of Tables for Statisticians and Biometricians, Cambridge University Press, 1914.

determine very easily the probability that a given value of χ^2 will be exceeded through random sampling errors alone.

In his formulation of the sampling distribution of χ^2, Pearson recognized, that to a limited extent, the calculated frequencies are forced to agree with the observed frequencies by the necessary requirement that Σc must equal Σo for comparability. The reader may readily grasp this point by returning to Table 35, empirically altering therein the values in the c column. He will find that, of the 7 values, he may change only 6 independently at will, the seventh always being determined by the requirement that Σc must equal 64 if the series of inscribed frequencies is to be comparable with the observed one. That is, only 6 of the 7 deviations $o - c$ are independent, for the seventh must be adjusted so that the total, $\Sigma(o - c)$, equals zero. Of the n' elements entering the sum called χ^2, only $n' - 1$ are then truly independent. Pearson's equation and Elderton's derived tables take account of this. It was not until nearly a quarter of a century later that Fisher[3] demonstrated that recognition of other restrictions limiting the number of independent elements should properly and may readily be made. Fisher pointed out that each independent statistic of the observed set of frequencies which is used in formulation of the calculated set imposes one restriction on the freedom of the c values to deviate from the o values. The number of independent deviations, n, in the χ^2 sum is the total number of elements, n', minus this number of restrictions, r.

$$n = n' - r. \tag{2}$$

The logic of this concept is unquestionably sound, although much acrimonious debate springing from conflicting definitions followed its introduction. Let the reader consider for a moment the distribution of sex at birth. If we assume that the probability of male sex is 0.5, then for Geissler's series we have:

Sex	Frequencies		$o-c$	$\frac{(o-c)^2}{c}$
	Observed	Calculated		
Males	2,582,914	2,508,816	74,098	2,188.5
Females	2,434,718	2,508,816	−74,098	2,188.5
Totals	5,017,632	5,017,632	0	$\chi^2 = 4,217.0$

[3] R. A. Fisher. On the interpretation of χ^2 from contingency tables, and the calculation of P. J. Royal Stat. Soc., 85: 87–94. 1922.

There is marked discordance between observation and theory, and χ^2 is not zero. However, the two elements $\frac{(o-c)^2}{c}$ in the sum are not independent. The $o - c$ values are numerically identical and opposite in sign, and this must of necessity be true. There is only one independent deviation, and the magnitude of χ^2 must be interpreted on this basis. By referring to appropriate tables, one finds that the probability of the discordance being due to chance is infinitesimal. What then is the true probability of male sex at birth? If one now uses $p = 0.514768$ derived from the observed data, then obviously c will agree with o in each case and there will be no discordance. χ^2 equals zero. Does this prove that the value $p = 0.514768$ is correct? Not at all! Theory has been forced to agree with observation here in two respects, $N_o = N_c$ and $p_o = p_c$, and any independent deviation is no longer possible.

The principle illustrated above in simple terms may be extended to all situations. The number of independent deviations contributing to the χ^2 sum is crucial to determination of the probability that the lack of complete concordance between the observed and theoretical frequencies might be due solely to errors of random sampling. This number of "independent deviations" is also designated as the number of "degrees of freedom," an expression that was current in the literature of statistical mechanics several decades ago.[4] Fisher's adoption of the "degrees of freedom" terminology has resulted in its wide dissemination in the statistical literature of the last decade. This writer suggests that the alternative term "independent deviations" is likely to be more comprehensible to those whose familiarity with mathematical theory is not very extensive. It will therefore be preferred in this discussion.

The sampling distribution of χ^2 with n independent deviations is defined by the equation

$$w = C(\chi^2)^{\frac{n-2}{2}} e^{-\frac{1}{2}\chi^2}, \tag{3}$$

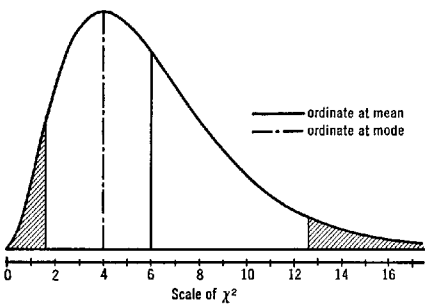
where C is a constant determined solely by n. This equation, established by Fisher,[3] is identical with the one derived earlier by Pearson,[1] subject to the condition that Pearson's n' be replaced by $n + 1$. Thus Elderton's tables [2] of the probability integral of χ^2 still hold, provided that n' therein be read as $n + 1$ rather than as the number of frequency classes.

[4] It is of interest to note that the underlying principle of adjusting for imposed restrictions was recognized in other connections at least as early as 1823. *Vide* C. F. Gauss. Theoria Combinationis Observationum Erroribus Minimis Obnoxiae, Pars posterior, Art. 38. Göttingen. 1823. (The author is indebted to Dr. W. Edwards Deming for this reference.)

The two shaded tails of the curve in Fig. 44 indicate by their truncating ordinates the values of χ^2 below and above which respectively only 5 per cent of values due to random sampling may be expected to occur when there are six independent deviations. Thus, values of χ^2 below 1.635 (for six independent deviations) are just as unlikely as values above 12.592. One is normally concerned solely with the latter type of value, however. It indicates a point above which values of χ^2 have so low a likelihood of occurrence solely through errors of sampling as to cause one to believe that, when they arise in practical work, the theoretical and observed sets of frequency really are not in good enough agreement to substantiate the theory. Elderton's table [2] gives to seven places of decimals the relative area in the tail beyond (greater than) successive values of χ^2 from zero to 70. We reproduce in Appendix IV to this volume a condensed table of χ^2 to three places of decimals only, and in terms of n, the number of independent deviations. When more detailed probabilities are desired, Elderton's table may be consulted, it being remembered that therein n' corresponds to our $n + 1$.

FIGURE 44

RANDOM SAMPLING DISTRIBUTION OF χ^2 WITH 6 INDEPENDENT DEVIATIONS



The mathematically adept reader will discern from equation (3) that the distribution of χ^2 is of skew form. Its mean is always at $\chi^2 = n$, the mode being two units less, that is at $\chi^2 = n - 2$. The curves for each value of n all start at $\chi^2 = 0$, and extend to infinite magnitudes. For small values of n the skewness is very marked. However, as n increases, the skewness gradually diminishes and the normal curve is approached as the limiting form of the χ^2 distribution. This feature is not readily demonstrable with a diagram drawn on the χ^2 scale because the visible range becomes very extended as n increases, with consequent fall in the

FIGURE 45

RANDOM SAMPLING DISTRIBUTIONS OF χ^2 FOR SEVERAL SMALL VALUES OF n *

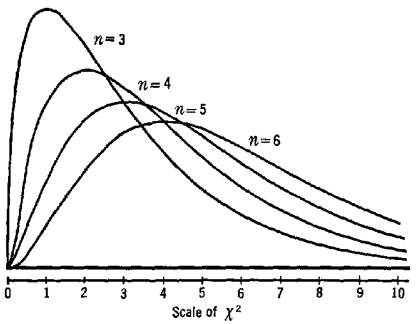
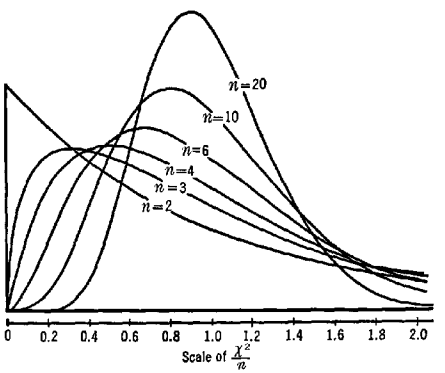


FIGURE 46

RANDOM SAMPLING DISTRIBUTIONS OF $\frac{\chi^2}{n}$ FOR SEVERAL VALUES OF n *



* These curves are reproduced from drawings very kindly supplied by Dr. W. Edwards Deming.

ordinate values in general. This is demonstrated in Fig. 45, wherein only four consecutive values of n are employed. The difficulty may be circumvented by transforming to the scale of $\frac{\chi^2}{n}$. Figure 46 presents a series of curves so drawn. The transition towards normality as n increases may be appreciated more easily from study of this diagram. In practical work, however, n is rarely large enough for the normal curve to provide an adequate approximation to the random sampling distribution of χ^2.

Figures 45 and 46 may, nevertheless, be studied advantageously. They serve to illustrate the most important point that the probability of any particular value of χ^2 being exceeded is highly dependent on the number of independent deviations embodied in its magnitude.

ILLUSTRATIONS

The practical application of the χ^2 test is very simple once the theoretical frequencies are evaluated. For the first two of the following illustrations, the theoretical frequencies are secured from the binomial expansion. In the third and fourth examples the theoretical frequencies are defined by continuous frequency curves. The reader will find in Table 24 another set of data to which he may apply the χ^2 criterion as a test of his command of the reasoning involved.

Example 1. An illustration in Chapter 12 of the use of the binomial expansion dealt with an experimental situation in which seeds of Pima cotton were planted 6 to a hill in 1,120 hills. Two weeks later a detailed count revealed that the surviving crop consisted of 3,320 seedlings, distributed in groups as given under o in Table 36. The adjoining calculated values c in that table are those derived by the binomial expansion and are copied directly from Table 21.[5] The lack of agreement between the observed frequencies and those given by the binomial expansion may be measured in terms of χ^2. Is χ^2 sufficiently large to warrant the lack of agreement being designated as "significant," and what does "significant" mean here?

The calculations in Table 36 reveal a very large value of χ^2 in this case. Though there are 7 classes in the sum, it contains but 5 independent elements, for the c distribution has been forced to agree with the o distribution with respect to two parameters, N and π. A glance under $n = 5$ in the table forming Appendix IV to this volume is sufficient to indicate that, with χ^2 so large, the probability of the discordance in frequency being due to chance must be essentially infinitesimal.

Study of the $o-c$ values in Table 36 reveals that the seedlings were more abundant in the highest and lowest survival classes, and correspondingly deficient in the intervening classes, than the independence requirement of the binomial permits.

[5] *Vide* page 175.

Thus they tended to survive or die as groups rather than as individual plants. Such a situation might arise through physical obstacles like patchy hard soil-crust preventing emergence of the seedlings as groups, or through the presence of diseases tending to take all seedlings in the groups they infect. The statistical analysis in terms of χ^2 merely signifies, of course, that the fates of seedlings within the groups are far from being independent.

TABLE 36

THE χ^2 CRITERION USED AS A TEST OF INDIVIDUAL INDEPENDENCE IN SURVIVORSHIP OF GROUPED PLANTS

x = number of surviving seedlings from 6 seeds per hill.

x	Frequencies		Discrepancies $o-c$	$\frac{(o-c)^2}{c}$
	o	c		
0	262	22.4	+239.6	2562.86
1	83	123.4	− 40.4	13.23
2	99	283.8	−184.8	120.33
3	164	348.2	−184.2	97.44
4	217	240.3	− 23.3	2.26
5	191	88.4	+102.6	119.08
6	104	13.6	+ 90.4	600.89
Totals	1,120	1120.1	0.0	χ^2=3516.09
$n = 5$. P_{χ^2} is infinitesimally small.				

Example 2. In an effort to demonstrate the applicability of the binomial theorem in forecasting the results of experiments, Weldon tossed 12 dice 26,306 times with thorough randomization between each throw. The appearance of either the 5 or the 6 face was recorded as a favorable event, π thus being established on *a priori* reasoning as one-third. The second and third columns of Table 37 present the observed and calculated frequency distributions of number of favorable events per throw. Are the experimental results in accord with theoretical expectancy?

There are 13 possible combinations of favorable and unfavorable events in this problem. The expected frequencies for 10, 11, and 12 favorable events are, however, quite small. These classes have been grouped in Table 37 because the random sampling distribution of χ^2 involves an assumption that the expected frequency shall always be large enough to give a reasonably normal distribution of its sampling error. This is undoubtedly achieved when c is 20 or above, but one will not err far if c is around 10 or more. Grouping the last 3 classes brings c to 14.3, which will be taken as an acceptable value. n' is reduced to 11 by this grouping, and the one restriction of equal total frequencies makes n equal to 10. With χ^2 equal to 35.6, P may be found from Elderton's table (entered under $n' = n + 1 = 11$) to be

0.00054. There is therefore only about 1 chance in 2,000 that so large a discrepancy would arise through errors of random sampling alone.

TABLE 37

APPLICATION OF THE χ^2 CRITERION TO WELDON'S DICE-TOSSING EXPERIMENT. THEORETICAL FREQUENCIES BASED ON THE BINOMIAL EXPANSION USING $\pi = \frac{1}{3}$

x = number of favorable results per toss.

x	Frequencies		Discrepancies $o-c$	$\frac{(o-c)^2}{c}$
	o	c		
0	185	202.8	− 17.8	1.56
1	1,149	1,216.5	− 67.5	3.75
2	3,265	3,345.4	− 80.4	1.93
3	5,475	5,575.6	−100.6	1.82
4	6,114	6,272.6	−158.6	4.01
5	5,194	5,018.0	176.0	6.17
6	3,067	2,927.2	139.8	6.68
7	1,331	1,254.5	76.5	4.67
8	403	392.0	11.0	.30
9	105	87.1	17.9	3.68
10	14			
11	4 } 18	14.3	3.7	.96
12	0			
Totals	26,306	26,306.0	0.0	$\chi^2 = 35.53$
$n = 10.$ $P_{\chi^2} = 0.00054.$				

Lack of independence in such a carefully executed random tossing experiment is not likely. One is accordingly disposed to explain the discrepancies between o and c by assuming that the dice were biased. This may be tested by calculating the probability p of the favorable event from the experimental throws. From Table 37,

$$p = \frac{\Sigma ox}{12(26{,}306)} = \frac{106{,}602}{315{,}672} = 0.3377,$$

a value decidedly in excess of that of one-third given by the assumption of unbiased dice. The statistical significance of the deviation, $p - \pi$, may be tested by applying the technique given in the previous chapter. For $\pi = \frac{1}{3}$ and $N = 315{,}672$,

$$\sigma_p = \sqrt{\frac{1}{3} \times \frac{2}{3} \times \frac{1}{315{,}672}} = 0.000839.$$

Then

$$k = \frac{p - \pi}{\sigma_p} = 5.2,$$

for which the corresponding P_k is approximatley 1 in 5,000,000. There is no doubt then that Weldon's dice must have been biased.

The technique of tossing may now be tested for its randomization of throws by determining a new theoretical set of frequencies having the same probability of the favorable event as the observed set. Those frequencies, together with the calculations for χ^2, are given in Table 38. There being two restrictions this time, n is

TABLE 38

THE χ^2 CRITERION USED AS A TEST OF INDIVIDUAL INDEPENDENCE OF DICE IN WELDON'S EXPERIMENT

$\pi = 0.3376986 = p$ observed.

x = number of favorable events per toss.

x	Frequencies		Discrepancies $o-c$	$\frac{(o-c)^2}{c}$
	o	c		
0	185	187.4	−2.4	0.03
1	1,149	1,146.5	2.5	0.01
2	3,265	3,215.2	49.8	0.77
3	5,475	5,464.7	10.3	0.02
4	6,114	6,269.4	−155.4	3.85
5	5,194	5,114.7	79.3	1.23
6	3,067	3,042.5	24.5	0.20
7	1,331	1,329.7	1.3	0.00
8	403	423.8	−20.8	1.02
9	105	96.0	9.0	0.84
10	14			
11	4 } 18	16.1	1.9	0.22
12	0			
Totals	26,306	26,306.0	0.0	$\chi^2=8.19$
$n = 9.$ $P_{\chi^2} = 0.52.$				

equal to 9. For χ^2 equal to 8.19, the probability of the frequency discordance being due to chance is seen from Appendix IV to be slightly in excess of 0.5, indicating excellent agreement. Thus Weldon's tossing was good. Probably nearly all dice are biased, as material removed in engraving the "dots" is seldom replaced by an equal amount of substance of the same specific gravity.

Example 3. An experiment to demonstrate the sampling distribution of means has been reported in Chapter 10. Therein, Fig. 31[6] shows the theoretical normal

[6] *Vide* page 132.

curve superimposed on the histogram for the means of 786 samples of 4 individuals each randomly drawn from a defined supply. The frequencies within specified ranges for both histogram and curve are given in Table 39. Is the histogram in agreement with the theoretically prescribed curve?

Only one of the three parameters defining the normal curve was drawn in this case from the observed frequencies. This restriction exists in every χ^2 test, for the

TABLE 39

TEST OF GOODNESS OF FIT OF THE THEORETICAL NORMAL CURVE TO THE DISTRIBUTION OF MEANS FROM 786 RANDOM SAMPLES OF 4.

Class range	Frequencies		Discrepancies $o-c$	$\frac{(o-c)^2}{c}$
	o	c		
→ 10.85	7 } 16	4.3 } 11.5	4.5	1.76
10.85–10.95	9 }	7.2 }		
10.95–11.05	21	15.9	5.1	1.64
10.05–11.15	17	30.5	−13.5	5.98
11.15–11.25	55	51.4	3.6	0.25
11.25–11.35	75	76.0	− 1.0	0.01
11.35–11.45	101	98.6	2.4	0.06
11.45–11.55	123	112.0	11.0	1.08
11.55–11.65	107	111.6	− 4.6	0.19
11.65–11.75	88	97.5	− 9.5	0.93
11.75–11.85	88	74.7	13.3	2.37
11.85–11.95	41	50.2	− 9.2	1.69
11.95–12.05	30	29.8	0.2	0.00
12.05–12.15	14	15.3	− 1.3	0.11
12.15–12.25	6 } 10	6.9 } 11.0	− 1.0	0.09
12.25→	4 }	4.1 }		
Totals	786	786.0	0.0	$\chi^2=16.16$
$n = 13.$ $P_{\chi^2} = 0.30.$				

total observed and calculated frequencies must agree if the test is to be applied. The mean and standard deviation of the curve are those theoretically prescribed; they are not derived from the observed distribution. The derived χ^2 of 16.16 therefore has $n = 13$ independent elements, and from Appendix IV one finds $P = 0.30$. The theory is therefore well substantiated by the outcome of this sampling experiment.

Example 4. The 500 determinations by Birge[7] of the width of a spectral band of light have been graduated by a normal curve in Fig. 1[8]. The classified frequencies

[7] These data were kindly supplied to the author by Professor Birge.

[8] *Vide* page 13.

are given in Table 40. Does the random error of measurement in determining the width of the light band appear to follow the normal law?

TABLE 40

TEST OF GOODNESS OF FIT OF THE NORMAL LAW TO THE DISTRIBUTION OF ERROR IN MEASURING WIDTH OF A SPECTRAL BAND OF LIGHT

x = measured width of band in microns.

x	Frequencies		Discrepancies $o-c$	$\frac{(o-c)^2}{c}$
	o	c		
→1705	5	4.7	−1.3	0.09
1705–1725	12 (17)	13.6 (18.3)		
1725–1745	43	36.1	6.9	1.32
1745–1765	61	70.6	−9.6	1.31
1765–1785	105	101.8	3.2	0.10
1785–1805	103	108.5	−5.5	0.28
1805–1825	89	85 3	3.7	0.16
1825–1845	54	49.5	4.5	0.41
1845–1865	19	21.3	−2.3	0.25
1865–1885	7	6.7	0.4	0.02
1885→	2 (9)	1.9 (8.6)		
Totals	500	500.0	0.0	$\chi^2=3.94$
$, n = 6.$ $P_{\chi^2}=0.68.$				

The question to be answered here deals solely with the form of a distribution and is not concerned with the values of the parameters defining the theoretical curve. One must logically use the curve of "best fit," that is, the one whose definitive parameters are derived from properties of the observed data. When fitting a normal curve by moments, this involves setting up the following identities:

	SAMPLE STATISTIC		CURVE PARAMETER
Frequency.....	N	=	N
Mean.....	$\bar{x}$	=	μ'
Standard deviation..	s_x	=	σ_x

Thus the curve used is forced to agree with the observed data in 3 respects, leaving $n = n' - 3$ independent elements in the χ^2 test. Since n' is 9 in Table 40, then the number of independent deviations left in χ^2 is 6. Now the derived χ^2 equals 3.94, and one may ascertain from Appendix IV that the probability of equal or greater departure from normality arising solely through chance is 68 in 100. The theory of normal distribution of error in this case is very well substantiated by the measurements made.

The graduation by a normal curve of the observed data on examination scores forming the histogram of Fig. 2*b*[9] may be tested as above. The data and calculations of that test are given in Table 41 because of the special interest of the resulting probability, $P_{\chi^2} = 0.98$. The graduation given by the normal curve is most unusually good in this case. Only once in 50 times would a better fit arise in a random sample. Is it "too good," thus placing the hypothesis of a normal distribution of the

TABLE 41

TEST OF GOODNESS OF FIT OF THE NORMAL LAW TO THE DISTRIBUTION OF INDIVIDUAL SCORES ON A "CURRENT AFFAIRS" EXAMINATION

Score on test	Frequencies		Discrepancies $o-c$	$\frac{(o-c)^2}{c}$
	o	c		
→49	2	2.9	0.7	0.05
50– 54	7 (9)	6.8 (9.7)		
55– 59	18	17.3	0.7	0.03
60– 64	36	35.4	0.6	0.01
65– 69	59	58.3	0.7	0.01
70– 74	74	77.3	−3.3	0.14
75– 79	88	82.9	5.1	0.31
80– 84	72	71.6	0.4	0.00
85– 89	44	49.7	−5.7	0.65
90– 94	30	27.9	2.1	0.16
95– 99	11	12.6		
100–104	5 (19)	4.6 (18.9)	0.1	0.00
105→	3	1.7		
Totals	449	449.0	0.0	$\chi^2 = 1.36$

$n = 7$. $P_{\chi^2} = 0.98$.

scores in jeopardy? It is true that this experience deviates from average expectancy of $P_{\chi^2} = 0.50$ just as much as a probability of 0.02, which might well be taken to designate a "significant deviation." Does it also signify that something is wrong? There are some who would argue so, but the burden rests with them to suggest on logical grounds wherein the agreement has been forced. In this specific example there are no grounds for such argument. It must be remembered that such probabilities do arise through chance, just as sometimes those below 0.02 do. One may

[9] *Vide* page 14.

well "raise the eyebrow" at statistical demonstrations that look too good, but in so doing the skeptic should not neglect the responsibility so assumed. The point should not be overlooked, however, that, when P_{χ^2} is extremely high, consideration of the *possibility* that factors forcing concordance may be operative is logically in order. The χ^2 test may justify examination of the possibility, but it does not logically signify forced concordance in the same sense that a low value of P_{χ^2} signifies lack of satisfactory agreement. If lack of satisfactory agreement were not *a priori* a logical alternative hypothesis to that of good agreement, then there would have been no point in testing the goodness of agreement at all.

CHAPTER 16

INDEPENDENCE AND BIVARIATE TABLES OF FREQUENCY

The reader may have noticed that, although the χ^2 criterion was applied in the preceding chapter solely to testing frequency concordance in univariate distributions of quantitatively described variables, the reasoning leading to its elaboration is not so restricted. χ^2 may be applied wherever observed frequencies are to be tested for concordance with a theoretical set applying to the same classes, the number of independent deviations being ascertainable. The classes providing the frequencies need not be quantitatively defined; division of the total frequency into classes representing types on some scale of description is all that is necessary. The concordance of theoretical Mendelian ratios with the outcome of hybridization experiments may be tested with confidence by means of this criterion, provided that the expected frequencies in crucial classes are not too small.

Application of the χ^2 technique to bivariate and multivariate frequency tables to test the independence from one another of the variables involved is very simple. It is commonly so employed when the variables are described through qualitative classification only. The definition of independence previously given[1] automatically provides the basis for computation of the necessary theoretical set of frequencies. These possibilities will now be investigated in terms of actual data.

The mouths of 135 women who claimed to have good health (aside from dental imperfections) were carefully examined to determine the periodontal condition[2]. Each mouth was classified with respect to degree of involvement with pyorrhea according to the arbitrary scale: *A* absent, *B* slight, *C* moderately advanced, *D* advanced and far advanced. Careful records were then made of the patients' diets during a so-called average week, each diet being freely chosen. These diets were later evaluated for the daily intake of the supposedly important mineral elements and vitamins. Formation of bivariate frequency distribution

[1] *Vide* page 170.

[2] The author is indebted to Dr. Dorothea Radusch of the School of Dentistry, University of Minnesota, for these data.

tables for the association of periodontal condition and daily intake of minerals and vitamins was thus made possible. Table 42 presents the data for daily calcium intake as it was found associated with periodontal condition. It is desired to determine whether there is any relationship between these two variables.

TABLE 42

OBSERVED DISTRIBUTION OF PERIODONTAL CONDITION AND AVERAGE DAILY CALCIUM INTAKE IN 135 WOMEN

			Periodontal condition				Totals
			A	*B*	*C*	*D*	
Calcium per day, in grams	0.70→	*a*	11	6	6	2	25
	0.55–0.70	*b*	10	8	3	1	22
	0.40–0.55	*c*	3	5	11	11	30
	→0.40	*d*	5	4	26	23	58
Totals			29	23	46	37	135

It is rather difficult to sense any correlation between the two variables in Table 42 through inspection of its frequencies. The necessarily categorical description of periodontal condition forbids appeal to the form of correlation analysis previously discussed. However, one may appeal to the definition that, if two events are independent, the probability of their joint occurrence is equal to the product of the probabilities of each occurring alone. If periodontal condition and calcium intake are independent of one another, then the 29 people having no pyorrhea (periodontal type *A*) should not differ in their calcium intake from those of any other type. They should be distributed in the given classes of calcium intake in the same proportions as each of the other four types. That is, the individuals in each periodontal group should be distributed with respect to calcium intake in the same proportions as are given by the group as a whole. The expected frequency in each cell, on the assumption of *independence* of the two variables, is determinable directly from the marginal totals alone.

The calculation of these expected frequencies is very simple. The 29 individuals of periodontal condition A are to be distributed among the four calcium intake classes, a to d, according to the respective proportions $\frac{25}{135}$, $\frac{22}{135}$, $\frac{30}{135}$, and $\frac{58}{135}$. Multiplying each proportion by the total 29 gives the required number. That is, on the assumption of independence, the expected frequency in each cell is one-Nth part of the product of the marginal totals for the two arrays[3] in which the cell lies. The calculation of χ^2 may then be proceeded with directly, it being convenient usually to prepare a table large enough to permit of the entries being made in each cell to which they belong. This is done in Table 43, wherein χ^2 is found to be 44.6.

An important problem is to determine the number of independent deviations of observation from theory in such tests of concordance. Since each array in the table of theoretical frequencies is derived by proportioning the observed total for that array, the expected frequency in one cell of the array is fixed entirely by what has been assigned to the others. If a certain total is to be divided into r parts, then only $r - 1$ parts may be written down independently or with complete freedom. The value of the remaining part is determined by the restriction that the total of the r parts must be a specified value. In the bivariate type of table this restriction applies to both the r vertical and the s horizontal arrays, leaving only $(r - 1)(s - 1)$ cells with unrestricted freedom to show deviation. The general rule for determining the number n of independent elements in χ^2 as calculated above for bivariate tables is therefore very simple. For Table 43, $n = (r - 1)(s - 1) = 3 \times 3 = 9$.

The probability of a χ^2 value of 44.6 with 9 "degrees of freedom" arising by sampling errors alone may be determined from Elderton's table to be less than 1 in 100,000. There will be no question then of the validity of the conclusion that periodontal condition in these women is not independent of the calcium intake in their freely chosen diets. One may next ask: What is the nature of the association? This may be seen readily by inspection of the differences $o - c$ in Table 43, wherein the excess or deficiency of the observed frequency with respect to the expected value, assuming independence, is given respectively in boldface and italic type. It will be observed quickly that good periodontal condition runs more freely with high calcium, and advanced pyorrhea runs more freely with low calcium, than the assumption of independence allows for.

[3] The term array is used herein for both vertical and horizontal lines of frequencies.

The reader may have noticed that the calculated frequencies of Table 43 are for the most part less than 10. The probability derived from χ^2

TABLE 43

THE χ^2 TEST OF INDEPENDENCE APPLIED TO THE DATA OF TABLE 42

			Periodontal condition				Totals
			A	*B*	*C*	*D*	
Calcium per day	*a*	*o*	11	6	6	2	25
		c	5.4	4.3	8.5	6.9	0.18519
		o−*c*	+5.6	+1.7	−2.5	−4.9	
		$\frac{(o-c)^2}{c}$	5.8	0.7	0.7	3.5	
	b	*o*	10	8	3	1	22
		c	4.7	3.7	7.5	6.0	0.16296
		o−*c*	+5.3	+4.3	−4.5	−5.0	
		$\frac{(o-c)^2}{c}$	6.0	5.0	2.7	4.2	
	c	*o*	3	5	11	11	30
		c	6.4	5.1	10.2	8.2	0.22222
		o−*c*	−3.4	−0.1	+0.8	+2.8	
		$\frac{(o-c)^2}{c}$	1.8	0.0	0.1	1.0	
	d	*o*	5	4	26	23	58
		c	12.5	9.9	19.8	15.9	0.42963
		o−*c*	−7.5	−5.9	+6.2	+7.1	
		$\frac{(o-c)^2}{c}$	4.5	3.5	1.9	3.2	
Totals			29	23	46	37	135
$\chi^2 = 44.6.$ $n = 9.$ $P_{\chi^2} < 10^{-5}.$							

may then be somewhat misleading. Coarser grouping may be resorted to in order to avoid this type of error. The classes *A* and *B*, *a* and *b* of Table 43 are united in Table 44, giving a new (3 × 3)fold classification.

The new c values will, like the o values, be the sum of those in the original cells which are now fused. The new $\frac{(o-c)^2}{c}$ elements will not,

TABLE 44

CALCULATION OF χ^2 FROM A COARSER CLASSIFICATION OF THE DATA OF TABLE 43

			Periodontal condition			Totals
			$A+B$	C	D	
Calcium per day	$a+b$	o	35	9	3	47
		c	18.1	16.0	12.9	
		$o-c$	+16.9	−7.0	−9.9	
		$\frac{(o-c)^2}{c}$	15.8	3.1	7.6	
	c	o	8	11	11	30
		c	11.5	10.2	8.2	
		$o-c$	−3.5	+0.8	+2.8	
		$\frac{(o-c)^2}{c}$	1.1	0.1	1.0	
	d	o	9	26	23	58
		c	22.4	19.8	15.9	
		$o-c$	−13.4	+6.2	+7.1	
		$\frac{(o-c)^2}{c}$	8.0	1.9	3.2	
Totals			52	46	37	135

$\chi^2 = 41.8. \quad n = 4. \quad P_{\chi^2} < 10^{-5}$

however, be the sums of those from the united cells. For the new classification, χ^2 falls to 41.8. Since n is now reduced to 4, P becomes less than 1 in 1,000,000, a value strengthening the previously formed conclusion.

A SHORT-CUT METHOD

There are somewhat quicker methods of calculating χ^2 for bivariate frequency tables wherein a test of the independence of the two variables is to be made. It may be observed that, in general,

$$\chi^2 = \Sigma\left[\frac{(o-c)^2}{c}\right]$$

$$= \Sigma\left[\frac{o^2}{c}\right] - 2\Sigma o + \Sigma c$$

$$= \Sigma\left[\frac{o^2}{c}\right] - N$$

$$= N\left(\Sigma\left[\frac{o^2}{Nc}\right] - 1\right).$$

It has been shown above that, in bivariate tables, c is one-Nth part of the product of the totals of the two arrays in which each cell lies. Therefore, if T_x and T_y are those totals, then

$$Nc = T_x T_y$$

and

$$\frac{o^2}{Nc} = \frac{o^2}{T_x T_y} = u, \text{ say.} \tag{1}$$

It is numerically simpler to calculate the values u than the elements $\frac{(o-c)^2}{c}$ in practical work, and therefore the formula

$$\chi^2 = N[\Sigma u - 1] \tag{2}$$

is often used. The disadvantage of this method lies in the c and $o - c$ values for each cell not being determined. One is therefore at a loss to know whether the grouping is too fine over any region. Also one cannot inspect the arrangement of the signs of $o - c$ in order to determine the nature of any association indicated by χ^2 being significant.

For illustrative purposes this method is applied in Table 45, where the relationship of periodontal condition to average daily intake of vitamin C is considered. Broad grouping of the original data is necessary to secure sufficiently large theoretical frequencies for this material. The values of c, which do not appear in these calculations, had to be derived independently for this purpose. χ^2 is very small this time, and with n equal to 2 the probability of the observed frequencies arising through random sampling from an uncorrelated surface is 0.24. There

is very good agreement therefore between the observed *facts and the* hypothesis that pyorrhea and vitamin C intake are independent of one another for the patients considered. With diets showing greater deficiency in vitamin C intake than are represented above, there might prove to be significant association with periodontal condition.

TABLE 45

EXTENT OF PYORRHEA AND THE AVERAGE DAILY VITAMIN C INTAKE IN 135 WOMEN

			Periodontal condition, x			Totals, T_y
			$A + B$	C	D	
Sherman units of vitamin C per day, y	60 and above	o	34	23	19	76
		T_xT_y	3952	3496	2812	
		u	0.2925	0.1513	0.1284	
	Less than 60	o	18	23	18	59
		T_xT_y	3068	2714	2183	
		u	0.1056	0.1949	0.1484	
Totals, T_x			52	46	37	135
$\chi^2 = N[\Sigma u - 1] = 2.85.$ $n = 2.$ $P = 0.24.$						

THE FOURFOLD TABLE

The situation is not at all uncommon, particularly in medical and sociological investigations, wherein characters under analysis may be divisible into no more than two types. In other inquiries it is a matter of expediency to coarsen the grouping in a more finely divisible variable so that only two contrasting classes are set up. Situations arise therefore in which the relationship between two variable characters are defined in a bivariate frequency table which has but four cells. Such tables are commonly known as "fourfold tables." Sex, marital status, survival, success, existence, maturity, etc., represent familiar conditions with respect to which description may be confined to two alternative classes.

For purposes of generalization one may designate such classifications as of a "presence-absence" type. The fourfold classification in joint

distribution of such variables may be portrayed in the general form of Table 46, wherein the letters designate the observed frequencies. The simplicity of this table naturally leads to simplifications of procedure in the computation of χ^2 to determine independence of the two characters. Let a', b', c', and d' indicate the expected cell frequencies on the assumption of independence. Then

$$a' = \frac{e \times g}{N},$$

and $a - a'$ is quickly determinable. Note, however, that $b - b'$ must equal $a - a'$ numerically but have the opposite sign; also $c - c'$ must equal $b - b'$, and $d - d'$ must equal $a - a'$. Thus only one expected frequency need be calculated in the usual way, the other three being written down by addition or subtraction of $a - a'$ from the observed frequency. This leads to a very quick computation of χ^2 using the basic formula.

TABLE 46

THE FOURFOLD TABLE IN GENERAL SYMBOLS

		Character x		Totals
		Present	Absent	
Character y	Present	a	b	e
	Absent	c	d	f
Totals.....		g	h	N

Following the general rule previously given, the fourfold table has but one "degree of freedom." It has just been noted that only one expected frequency need be determined from the array totals; the others are dependent on the deviation of the theoretical frequency from the observed in this one cell. Now the sampling distribution of χ^2 when n equals unity is very closely related to the normal curve. Indeed, if the square root of χ^2 is taken, then the sampling distribution of χ thus secured is the positive half of the normal curve with χ equal to the relative deviate k. Since 1.96 is the 5 per cent point in k, then 3.84 is the 5 per cent point in the χ^2 distribution with but one independent element.

The adjacent table of observed frequencies for the effect of previous vaccination on recovery from smallpox in 2,081 cases of the disease has been provided by Pearson.[4] It is not surprising to find from the computations in Table 47 that χ^2 is so large (177.3) that the probability of independence of the two variables is negligible.

TABLE 47

EFFECT OF PRIOR VACCINATION ON RECOVERY FROM SMALLPOX

(Data from K. Pearson[4])

			Effect of smallpox		Totals
			Recovery	Death	
Vaccination cicatrix	Present	o	1,562	42	1,604
		c	1,499	105	
		$o-c$	+63	−63	
		$\frac{(o-c)^2}{c}$	2.6	37.8	
	Absent	o	383	94	477
		c	446	31	
		$o-c$	−63	+63	
		$\frac{(o-c)^2}{c}$	8.9	128.0	
		Totals	1,945	136	2,081

$\chi^2 = 177.3.$ $n = 1.$ $P_{\chi^2} < 10^{-6}$

THE FOURFOLD TABLE WITH SMALL FREQUENCIES

Let us assume now that Pearson had available only about one-tenth as many cases, distributed in essentially the same proportions. The distribution in Table 48 is suggested. Would demonstration of the lack of independence of the two variables still hold?

[4] Drapers' Company Research Memoirs, Biometric Series VII. London. 1912.

The χ^2 criterion may be applied as before, but it is worthy of note that, when the cell frequencies a, b, c, and d, of the fourfold table are

TABLE 48

SHORT-CUT CALCULATION OF χ^2 USING EQUATION (3)

Imaginary data derived from Table 47

		Effect of smallpox		Totals
		Recovery	Death	
Vaccination cicatrix	Present	156	4	160
	Absent	38	9	47
Totals		194	13	207

$$\chi^2 = \frac{324{,}473{,}228}{18{,}965{,}440} = 17.1 \quad n = 1. \quad P_{\chi^2} = 0.00004.$$

small, a short-cut formula for χ^2 may be applied with some saving of time. The alternative formula is:

$$\chi^2 = \frac{N(ad - bc)^2}{e \cdot f \cdot g \cdot h}. \tag{3}$$

This equation is very easy to remember if the following points are noted:

(1) ad and bc are the products of the diagonally opposed cells of the table, and

(2) the denominator is the product of the 4 array totals derived from these cells.

It is only a matter of routine elementary algebra, which need not be given here, to show the identity of this equation with the fundamental one applied to a fourfold table. Using equation (3) in Table 48 because the cell frequencies are not very large, one finds χ^2 equals 17.1, a value clearly significant of association. The answer could have been given, however, without further calculation, for, with a given set of *proportions*, χ^2 varies in direct ratio with N.

$$\chi^2 = \Sigma\left[\frac{(o - c)^2}{c}\right]$$

$$= N\Sigma\left[\frac{\left(\frac{o}{N} - \frac{c}{N}\right)^2}{\frac{c}{N}}\right]$$

Had the frequencies been reduced to precisely one-tenth their previous values, χ^2 would have been reduced in like manner.

This is a very useful proposition to remember. With the same number of independent deviations in which the ratio of o to c is preserved but N is varied, χ^2 will be directly proportional to N. That is, as N increases in such a situation, the probability of the deviations arising through sampling errors alone will fall. This is in accord with logical demands, of course. However, the point is too easily overlooked that concordance between an observed and a theoretical set of frequencies that may appear to be excellent in a geometric representation may nevertheless prove to be very poor by the χ^2 criterion if N is very large. And inversely, when N is small, the χ^2 criterion may be quite insensitive to a lack of concordance that, to the student of science, may appear to be important. In the problems of science, consideration of the importance of discrepancies should be guided, *not replaced*, by contemplation of statistical significance tests.

The question of a differential case fatality between the sexes in the Mankato typhoid epidemic of 1908 has been considered previously.[5] A test of the significance of the difference between the two proportions of fatality showed that the probability of the observed discrepancy arising through sampling errors was quite high ($k = 0.305$, $P = 0.76$). This calculation might equally well have been handled in terms of frequency instead of proportion, using the χ^2 criterion in place of the difference in rate test there given. Table 49 presents the calculations. One hardly needs proceed beyond securing the first expected cell frequency to realize that the concordance of observation with expectation on the theory of no sex difference is excellent. The full calculations are shown, however, to demonstrate the identity of the two tests as solutions to the same problem. Any pair of proportions or rates, derived from samples of specified sizes, may be tested for concordance in terms of the proportions or the frequencies, as desired. The answer will be the same, although the two sets of calculations seem widely different. According to this principle, heterogeneity in a set of more than two proportions or rates may be tested very simply by transforming to frequencies and using the χ^2 criterion.

[5] *Vide* page 206.

TABLE 49

TEST FOR INDEPENDENCE OF SEX AND MORTALITY FROM DISEASE. TYPHOID FEVER EPIDEMIC IN MANKATO, 1908

			Result of disease		Total cases
			Death	Survival	
Sex	Male	o	16	205	221
		c	15.137	205.863	
		$o-c$	+0.863	*−0.863*	
		$\frac{(o-c)^2}{c}$	0.0492	0.0036	
	Female	o	19	271	290
		c	19.863	270.137	
		$o-c$	*−0.863*	+0.863	
		$\frac{(o-c)^2}{c}$	0.0375	0.0028	
		Totals	35	476	511

$\chi^2 = 0.0931.$ $\chi = 0.305.$ $n = 1.$ $P_{\chi^2} = 0.76.$

APPENDIX I

A TABLE OF NORMAL CURVE FUNCTIONS

k is the relative deviate of the measure which is normally distributed. If x is the variable, then

$$k = \frac{x - \bar{x}}{s_x}, \quad \text{or} \quad k = \frac{x - \mu_x'}{\sigma_x}.$$

P is the probability of the k magnitude being exceeded solely through errors of random sampling.

$$P = \frac{\text{Area in tail beyond } k}{\text{Area of curve segment having same sign as } k}.$$

w is the ordinate at k for a curve of unit area and unit standard deviation.

The reader may note carefully the definition above of the tabled probability. Since the normal curve is symmetrical about $k = 0$, this probability is identical with that defined by the ratio of the area in *both* tails, ignoring the sign of k, to the area of the *whole* curve. It is in the latter form that the probability is usually defined. The sign of k is ordinarily considered of no consequence since it may be reversed at will. The choice between $\bar{x} - \bar{y}$ and $\bar{y} - \bar{x}$ may justly be regarded as a perfectly empirical one, and as a consequence the deviation alone may be considered important, not its sign. The simplicity of this type of reasoning has considerable appeal, but it is not without its difficulties when the sampling distribution to be used in testing the significance of the difference between other statistics proves to be skew. As a result of repeated reflection on this problem in its general form, the author prefers to consider difference problems according to either one of the patterns given below.

Let a and b represent two statistics, the significance of the difference between which is to be judged. Then two intimately related propositions arise, as follows:

(1) If the sign of $a - b$ is of consequence, then one is concerned with determining how often, among all possible differences *of the same sign* which will occur through errors of random sampling alone, one *as great as* $a - b$ will arise. This will involve considering a single tail of the difference sampling distribution in relation to the curve segment of the same sign.

(2) If the sign of $a - b$ is logically to be ignored, then $|a - b|$ defines a tail in a new distribution of such quantities as itself which would arise solely through errors of random sampling. This distribution will start at zero. Also, if the $a - b$ distribution is symmetrical, then the $|a - b|$ distribution is coincident in form with the positive half of the $a - b$ distribution, and the probability will be identical with that given under (1) above.

Since the normal curve is symmetrical, solutions for both propositions above become identical in any specific instance.

Table of Normal Curve Functions

k	P	w	k	P	w
.00	1	.399	.50	.6171	.352
.01	.9920	.399	.51	.6101	.350
.02	.9840	.399	.52	.6031	.348
.03	.9761	.399	.53	.5961	.347
.04	.9681	.399	.54	.5892	.345
.05	.9601	.398	.55	.5823	.343
.06	.9522	.398	.56	.5755	.341
.07	.9442	.398	.57	.5687	.339
.08	.9362	.398	.58	.5619	.337
.09	.9283	.397	.59	.5552	.335
.10	.9203	.397	.60	.5485	.333
.11	.9124	.397	.61	.5419	.331
.12	.9045	.396	.62	.5353	.329
.13	.8966	.396	.63	.5287	.327
.14	.8887	.395	.64	.5222	.325
.15	.8808	.394	.65	.5157	.323
.16	.8729	.394	.66	.5093	.321
.17	.8650	.393	.67	.5029	.319
.18	.8572	.393	.68	.4965	.317
.19	.8493	.392	.69	.4902	.314
.20	.8415	.391	.70	.4839	.312
.21	.8337	.390	.71	.4777	.310
.22	.8259	.389	.72	.4715	.308
.23	.8181	.389	.73	.4654	.306
.24	.8103	.388	.74	.4593	.303
.25	.8026	.387	.75	.4533	.301
.26	.7949	.386	.76	.4473	.299
.27	.7872	.385	.77	.4413	.297
.28	.7795	.384	.78	.4354	.294
.29	.7718	.383	.79	.4295	.292
.30	.7642	.381	.80	.4237	.290
.31	.7566	.380	.81	.4179	.287
.32	.7490	.379	.82	.4122	.285
.33	.7414	.378	.83	.4065	.283
.34	.7339	.377	.84	.4009	.280
.35	.7263	.375	.85	.3953	.278
.36	.7188	.374	.86	.3898	.276
.37	.7114	.373	.87	.3843	.273
.38	.7039	.371	.88	.3789	.271
.39	.6965	.370	.89	.3735	.268
.40	.6892	.368	.90	.3681	.266
.41	.6818	.367	.91	.3628	.264
.42	.6745	.365	.92	.3576	.261
.43	.6672	.364	.93	.3524	.259
.44	.6599	.362	.94	.3472	.256
.45	.6527	.361	.95	.3421	.254
.46	.6455	.359	.96	.3371	.252
.47	.6384	.357	.97	.3320	.249
.48	.6312	.356	.98	.3271	.247
.49	.6241	.354	.99	.3222	.244

Table of Normal Curve Functions

k	P	w	k	P	w
1.00	.3173	.242	1.50	.1336	.130
1.01	.3125	.240	1.51	.1310	.128
1.02	.3077	.237	1.52	.1285	.126
1.03	.3030	.235	1.53	.1260	.124
1.04	.2983	.232	1.54	.1236	.122
1.05	.2937	.230	1.55	.1211	.120
1.06	.2891	.227	1.56	.1188	.118
1.07	.2846	.225	1.57	.1164	.116
1.08	.2801	.223	1.58	.1141	.115
1.09	.2757	.220	1.59	.1118	.113
1.10	.2713	.218	1.60	.1096	.111
1.11	.2670	.215	1.61	.1074	.109
1.12	.2627	.213	1.62	.1052	.107
1.13	.2585	.211	1.63	.1031	.106
1.14	.2543	.208	1.64	.1010	.104
1.15	.2501	.206	1.65	.0989	.102
1.16	.2460	.204	1.66	.0969	.101
1.17	.2420	.201	1.67	.0949	.099
1.18	.2380	.199	1.68	.0930	.097
1.19	.2340	.197	1.69	.0910	.096
1.20	.2301	.194	1.70	.0891	.094
1.21	.2263	.192	1.71	.0873	.092
1.22	.2225	.190	1.72	.0854	.091
1.23	.2187	.187	1.73	.0836	.089
1.24	.2150	.185	1.74	.0819	.088
1.25	.2113	.183	1.75	.0801	.086
1.26	.2077	.180	1.76	.0784	.085
1.27	.2041	.178	1.77	.0767	.083
1.28	.2005	.176	1.78	.0751	.082
1.29	.1971	.174	1.79	.0735	.080
1.30	.1936	.171	1.80	.0719	.079
1.31	.1902	.169	1.81	.0703	.078
1.32	.1868	.167	1.82	.0688	.076
1.33	.1835	.165	1.83	.0672	.075
1.34	.1802	.163	1.84	.0658	.073
1.35	.1770	.160	1.85	.0643	.072
1.36	.1738	.158	1.86	.0629	.071
1.37	.1707	.156	1.87	.0615	.069
1.38	.1676	.154	1.88	.0601	.068
1.39	.1645	.152	1.89	.0588	.067
1.40	.1615	.150	1.90	.0574	.066
1.41	.1585	.148	1.91	.0561	.064
1.42	.1556	.146	1.92	.0549	.063
1.43	.1527	.144	1.93	.0536	.062
1.44	.1499	.141	1.94	.0524	.061
1.45	.1471	.139	1.95	.0512	.060
1.46	.1443	.137	1.96	.0500	.058
1.47	.1416	.135	1.97	.0488	.057
1.48	.1389	.133	1.98	.0477	.056
1.49	.1362	.131	1.99	.0466	.055

Table of Normal Curve Functions

k	P	w	k	P	w
2.00	.0455	.054	2.50	.0124	.018
2.01	.0444	.053	2.51	.0121	.017
2.02	.0434	.052	2.52	.0117	.017
2.03	.0424	.051	2.53	.0114	.016
2.04	.0414	.050	2.54	.0111	.016
2.05	.0404	.049	2.55	.0188	.015
2.06	.0394	.048	2.56	.0105	.015
2.07	.0385	.047	2.57	.0102	.015
2.08	.0375	.046	2.58	.0099	.014
2.09	.0366	.045	2.59	.0096	.014
2.10	.0357	.044	2.60	.0093	.014
2.11	.0349	.043	2.61	.0091	.013
2.12	.0340	.042	2.62	.0088	.013
2.13	.0332	.041	2.63	.0085	.013
2.14	.0324	.040	2.64	.0083	.012
2.15	.0316	.040	2.65	.0080	.012
2.16	.0308	.039	2.66	.0078	.012
2.17	.0300	.038	2.67	.0076	.011
2.18	.0293	.037	2.68	.0074	.011
2.19	.0285	.036	2.69	.0071	.011
2.20	.0278	.035	2.70	.0069	.010
2.21	.0271	.035	2.71	.0067	.010
2.22	.0264	.034	2.72	.0065	.010
2.23	.0257	.033	2.73	.0063	.010
2.24	.0251	.032	2.74	.0061	.009
2.25	.0244	.032	2.75	.0060	.009
2.26	.0238	.031	2.76	.0058	.009
2.27	.0232	.030	2.77	.0056	.009
2.28	.0226	.030	2.78	.0054	.008
2.29	.0220	.029	2.79	.0053	.008
2.30	.0214	.028	2.80	.0051	.008
2.31	.0209	.028	2.81	.0050	.008
2.32	.0203	.027	2.82	.0048	.007
2.33	.0198	.026	2.83	.0047	.007
2.34	.0193	.026	2.84	.0045	.007
2.35	.0188	.025	2.85	.0044	.007
2.36	.0183	.025	2.86	.0042	.007
2.37	.0178	.024	2.87	.0041	.006
2.38	.0173	.023	2.88	.0040	.006
2.39	.0168	.023	2.89	.0039	.006
2.40	.0164	.022	2.90	.0037	.006
2.41	.0160	.022	2.91	.0036	.006
2.42	.0155	.021	2.92	.0035	.006
2.43	.0151	.021	2.93	.0034	.005
2.44	.0147	.020	2.94	.0033	.005
2.45	.0143	.020	2.95	.0032	.005
2.46	.0139	.019	2.96	.0031	.005
2.47	.0135	.019	2.97	.0030	.005
2.48	.0131	.018	2.98	.0029	.005
2.49	.0128	.018	2.99	.0028	.005

Table of Normal Curve Functions

k	P	w	k	P	w
3.00	.0027	.004	3.50	.0005	.001
3.01	.0026	.004	3.51	.0004	.001
3.02	.0025	.004	3.52	.0004	.001
3.03	.0024	.004	3.53	.0004	.001
3.04	.0024	.004	3.54	.0004	.001
3.05	.0023	.004	3.55	.0004	.001
3.06	.0022	.004	3.56	.0004	.001
3.07	.0021	.004	3.57	.0004	.001
3.08	.0021	.003	3.58	.0003	.001
3.09	.0020	.003	3.59	.0003	.001
3.10	.0019	.003	3.60	.0003	.001
3.11	.0019	.003	3.61	.0003	.001
3.12	.0018	.003	3.62	.0003	.001
3.13	.0017	.003	3.63	.0003	.001
3.14	.0017	.003	3.64	.0003	.001
3.15	.0016	.003	3.65	.0003	.001
3.16	.0016	.003	3.66	.0003	.000
3.17	.0015	.003	3.67	.0002	.000
3.18	.0015	.003	3.68	.0002	.000
3.19	.0014	.002	3.69	.0002	.000
3.20	.0014	.002	3.70	.0002	.000
3.21	.0013	.002	3.71	.0002	.000
3.22	.0013	.002	3.72	.0002	.000
3.23	.0012	.002	3.73	.0002	.000
3.24	.0012	.002	3.74	.0002	.000
3.25	.0012	.002	3.75	.0002	.000
3.26	.0011	.002	3.76	.0002	.000
3.27	.0011	.002	3.77	.0002	.000
3.28	.0010	.002	3.78	.0002	.000
3.29	.0010	.002	3.79	.0002	.000
3.30	.0010	.002	3.80	.0001	.000
3.31	.0009	.002	3.81	.0001	.000
3.32	.0009	.002	3.82	.0001	.000
3.33	.0009	.002	3.83	.0001	.000
3.34	.0008	.002	3.84	.0001	.000
3.35	.0008	.001	3.85	.0001	.000
3.36	.0008	.001	3.86	.0001	.000
3.37	.0008	.001	3.87	.0001	.000
3.38	.0007	.001	3.88	.0001	.000
3.39	.0007	.001	3.89	.0001	.000
3.40	.0007	.001	3.90	.0001	.000
3.41	.0006	.001	3.91	.0001	.000
3.42	.0006	.001	3.92	.0001	.000
3.43	.0006	.001	3.93	.0001	.000
3.44	.0006	.001	3.94	.0001	.000
3.45	.0006	.001	3.95	.0001	.000
3.46	.0005	.001	3.96	.0001	.000
3.47	.0005	.001	3.97	.0001	.000
3.48	.0005	.001	3.98	.0001	.000
3.49	.0005	.001	3.99	.0001	.000

APPENDIX II

TABLES OF z_r AS A FUNCTION OF r

$$z_r = \tfrac{1}{2}[\log_e(1+r) - \log_e(1-r)] = 1.1513\left[\log\frac{1+r}{1-r}\right]$$

TABLE A. r FROM ZERO TO 0.89 BY HUNDREDTHS

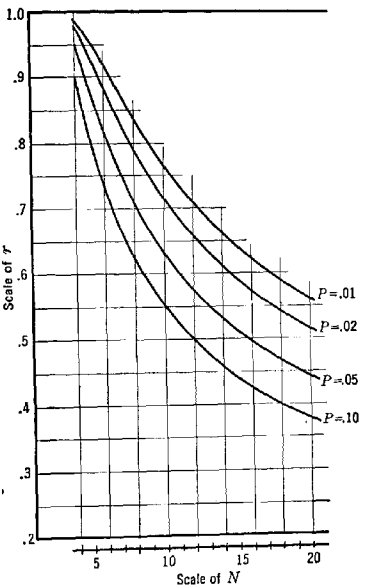
r	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
.0	.0000	.0100	.0200	.0300	.0400	.0500	.0601	.0701	.0802	.0902
.1	.1003	.1104	.1206	.1307	.1409	.1511	.1614	.1717	.1820	.1923
.2	.2027	.2132	.2237	.2342	.2448	.2554	.2661	.2769	.2877	.2986
.3	.3095	.3205	.3316	.3428	.3541	.3654	.3769	.3884	.4001	.4118
.4	.4236	.4356	.4477	.4599	.4722	.4847	.4973	.5101	.5230	.5361
.5	.5493	.5627	.5763	.5901	.6042	.6184	.6328	.6475	.6625	.6777
.6	.6931	.7089	.7250	.7414	.7582	.7753	.7928	.8107	.8291	.8480
.7	.8673	.8872	.9076	.9287	.9505	.9730	.9962	1.0203	1.0454	1.0714
.8	1.0986	1.1270	1.1568	1.1881	1.2212	1.2562	1.2933	1.3331	1.3758	1.4219

TABLE B. r FROM 0.900 TO 0.999 BY THOUSANDTHS

r	.000	.001	.002	.003	.004	.005	.006	.007	.008	.009
.90	1.4722	1.4775	1.4828	1.4882	1.4937	1.4992	1.5047	1.5103	1.5160	1.5217
.91	1.5275	1.5334	1.5393	1.5453	1.5513	1.5574	1.5636	1.5698	1.5762	1.5826
.92	1.5890	1.5956	1.6022	1.6089	1.6157	1.6226	1.6296	1.6366	1.6438	1.6510
.93	1.6584	1.6658	1.6734	1.6811	1.6888	1.6967	1.7047	1.7129	1.7211	1.7295
.94	1.7380	1.7467	1.7555	1.7645	1.7736	1.7828	1.7923	1.8019	1.8117	1.8216
.95	1.8318	1.8421	1.8527	1.8635	1.8745	1.8857	1.8972	1.9090	1.9210	1.9333
.96	1.9459	1.9588	1.9721	1.9857	1.9996	2.0139	2.0287	2.0439	2.0595	2.0756
.97	2.0923	2.1095	2.1273	2.1457	2.1649	2.1847	2.2054	2.2269	2.2494	2.2729
.98	2.2976	2.3235	2.3507	2.3796	2.4101	2.4427	2.4774	2.5147	2.5550	2.5987
.99	2.6467	2.6996	2.7587	2.8257	2.9031	2.9945	3.1063	3.2504	3.4534	3.8002

APPENDIX III

A GRAPH OF PROBABILITY LEVELS FOR THE RANDOM SAMPLING DISTRIBUTIONS OF r WHEN ρ IS EQUAL TO ZERO AND N IS SMALL



Examples of application

(1) With N equal to 20, a value $r = 0.55$ is obtained. By locating this point on the graph it may be seen that only once or twice in 100 trials would a value of r as high as the one observed arise through random sampling errors when ρ is zero.

(2) When $N = 5$, approximately 10 per cent of random samples from an uncorrelated supply will yield correlation coefficients exceeding 0.8.

APPENDIX IV

TABLE OF THE PROBABILITY THAT χ^2, DERIVED FROM n INDEPENDENT ELEMENTS, WILL BE EXCEEDED SOLELY THROUGH ERRORS OF RANDOM SAMPLING

Probability Integral of χ^2

χ^2	Number of independent elements, n									
	1	2	3	4	5	6	7	8	9	10
1	.317	.607	.801	.910	.963	.986	.995	.998	.999	.999
2	.157	.368	.572	.736	.849	.920	.960	.981	.991	.996
3	.083	.223	.392	.558	.700	.809	.885	.934	.964	.981
4	.046	.135	.261	.406	.549	.677	.780	.857	.911	.947
5	.025	.082	.172	.287	.416	.544	.660	.758	.834	.891
6	.014	.050	.112	.199	.306	.423	.540	.647	.740	.815
7	.008	.030	.072	.136	.221	.321	.429	.537	.637	.725
8	.005	.018	.046	.092	.156	.238	.333	.433	.534	.629
9	.003	.011	.029	.061	.109	.174	.253	.342	.437	.532
10	.002	.007	.019	.040	.075	.125	.189	.265	.350	.440
11	.001	.004	.012	.027	.051	.088	.139	.202	.276	.358
12	.001	.002	.007	.017	.035	.062	.101	.151	.213	.285
13	**	.002	.005	.011	.023	.043	.072	.112	.163	.224
14	**	.001	.003	.007	.016	.030	.051	.082	.122	.173
15	**	.001	.002	.005	.010	.020	.036	.059	.091	.132
16	**	**	.001	.003	.007	.014	.025	.042	.067	.100
17	**	**	.001	.002	.004	.009	.017	.030	.049	.074
18	**	**	**	.001	.003	.006	.012	.021	.035	.055
19	**	**	**	.001	.002	.004	.008	.015	.025	.040
20	**	**	**	**	.001	.003	.006	.010	.018	.029
21	**	**	**	**	.001	.002	.004	.007	.013	.021
22	**	**	**	**	.001	.001	.003	.005	.009	.015
23	**	**	**	**	**	.001	.002	.003	.006	.011
24	**	**	**	**	**	.001	.001	.002	.004	.008
25	**	**	**	**	**	**	.001	.002	.003	.005
26	**	**	**	**	**	**	.001	.001	.002	.004
27	**	**	**	**	**	**	**	.001	.001	.003
28	**	**	**	**	**	**	**	**	.001	.002
29	**	**	**	**	**	**	**	**	.001	.001
30	**	**	**	**	**	**	**	**	**	.001

** Less than .0005.

Probability Integral of χ^2

χ^2	Number of independent elements n									
	11	12	13	14	15	16	17	18	19	20
1	.999	.999	.999	.999	*	*	*	*	*	*
2	.998	.999	.999	.999	.999	.999	.999	.999	*	*
3	.991	.996	.998	.999	.999	.999	.999	.999	.999	.999
4	.970	.983	.991	.995	.998	.999	.999	.999	.999	.999
5	.931	.958	.975	.986	.992	.996	.998	.999	.999	.999
6	.873	.916	.946	.966	.980	.988	.993	.996	.998	.999
7	.799	.858	.902	.935	.958	.973	.984	.990	.994	.997
8	.713	.785	.844	.889	.924	.949	.967	.979	.987	.992
9	.622	.703	.773	.831	.878	.913	.940	.960	.973	.983
10	.530	.616	.694	.762	.820	.867	.904	.932	.953	.968
11	.443	.529	.611	.686	.753	.809	.857	.894	.924	.946
12	.363	.446	.528	.606	.679	.744	.800	.847	.886	.916
13	.293	.369	.448	.527	.602	.673	.736	.792	.839	.877
14	.233	.301	.374	.450	.526	.599	.667	.729	.784	.830
15	.182	.241	.307	.378	.451	.525	.595	.662	.723	.776
16	.141	.191	.249	.313	.382	.453	.524	.593	.657	.717
17	.108	.150	.199	.256	.319	.386	.454	.523	.590	.653
18	.082	.116	.158	.207	.263	.324	.389	.456	.522	.587
19	.061	.089	.123	.165	.214	.269	.329	.392	.457	.522
20	.045	.067	.095	.130	.172	.220	.274	.333	.395	.458
21	.033	.050	.073	.102	.137	.179	.226	.279	.337	.397
22	.024	.038	.055	.079	.108	.143	.185	.232	.284	.341
23	.018	.028	.042	.060	.084	.114	.149	.191	.237	.289
24	.013	.020	.031	.046	.065	.090	.119	.155	.196	.242
25	.009	.015	.023	.035	.050	.070	.095	.125	.161	.201
26	.006	.011	.017	.026	.038	.054	.074	.100	.130	.166
27	.005	.008	.012	.019	.029	.041	.058	.079	.105	.135
28	.003	.006	.009	.014	.022	.032	.045	.062	.083	.109
29	.002	.004	.007	.010	.016	.024	.035	.048	.066	.088
30	.002	.003	.005	.008	.012	.018	.026	.037	.052	.070

* Greater than .9995.

APPENDIX V

SELECTED FORMULAS

Definitions of Statistics

Arithmetic mean

$$\bar{x} = \frac{\Sigma x}{N}$$

Standard deviation

$$s_x = \sqrt{\frac{\Sigma(x - \bar{x})^2}{N}}$$

$$= \sqrt{\frac{\Sigma x^2}{N} - \bar{x}^2}$$

Variance

$$v_x = \frac{\Sigma(x - \bar{x})^2}{N - 1}$$

Maximum likelihood estimate of σ from s

$$\text{M.L.E.}_{\sigma_x} = \sqrt{v_x} = s_x \sqrt{\frac{N}{N - 1}}$$

Coefficient of variation

$$\text{C.V.}_x = \frac{100\, s_x}{\bar{x}}$$

Relative deviate

$$k_x = \frac{x - \bar{x}}{s_x}$$

Correlation coefficient

$$r_{xy} = \frac{\Sigma k_x k_y}{N}$$

$$= \frac{\dfrac{\Sigma xy}{N} - \bar{x}\bar{y}}{s_x s_y}$$

Regression coefficient

$$b_y = r_{xy}\frac{s_y}{s_x}$$

Rectilinear regression equation

$$\begin{aligned} Y &= a_y + b_y x \\ &= (\bar{y} - b_y\bar{x}) + b_y x \\ &= \left[\bar{y} - r_{xy}\frac{s_y}{s_x}\bar{x}\right] + \left[r_{xy}\frac{s_y}{s_x}\right]x \end{aligned}$$

Standard error of estimate, or standard deviation of the residual variation about the rectilinear regression line

$$s_{y.x} = s_y\sqrt{1 - r_{xy}^2}$$

Coefficient of alienation

$$\text{C.A.} = \sqrt{1 - r_{xy}^2}$$

Coding and decoding equations

If $\mathrm{x} = \dfrac{x - a}{b}$, a and b being selected suitable constants, then

$$x = a + b\mathrm{x}$$

$$\bar{x} = a + b\bar{\mathrm{x}}$$

$$s_x = bs_{\mathrm{x}}$$

and

$$r_{xy} = r_{\mathrm{xy}}$$

Moment coefficients. Quantities generating descriptions of the properties of frequency distributions

$$m_n = \frac{\Sigma(x - \bar{x})^n}{N} \qquad (n\text{th moment coefficient})$$

"Beta" and "Gamma" coefficients

$$b_1 = \frac{m_3^2}{m_2^3}, \quad \text{or} \quad g_1 = \frac{m_3}{s^3} \quad \text{(indices of skewness)}$$

$$b_2 = \frac{m_4}{m_2^2}, \quad \text{or} \quad g_2 = \frac{m_4}{m_2^2} - 3 \text{ (indices of kurtosis)}$$

(Names are derived from the equivalent Greek letters used for the supply parameters.)

χ^2, a statistical measure of lack of agreement between observed frequencies and theoretical expectancy in a complete set of n' classes, is defined in various ways. The general definition is

$$\chi^2 = \sum^{n'} \left[\frac{(o - c)^2}{c} \right]$$

In testing independence in bivariate frequency tables, this may be rearranged to

$$\chi^2 = N \left(\sum \left[\frac{o^2}{T_x T_y} \right] - 1 \right) \quad (\textit{vide} \text{ page } 232)$$

For the fourfold table it may be shown that

$$\chi^2 = \frac{N(ad - bc)^2}{e \cdot f \cdot g \cdot h} \quad (\textit{vide} \text{ Table 46, page } 234)$$

Frequency Distributions

The normal curve. A frequency function defined by the equation

$$w = Ce^{-\frac{1}{2}k^2}$$

where w = the ordinate at k

$$C = \frac{N}{s\sqrt{2\pi}}$$

$\pi = 3.1416 \cdots$

e = the natural base of logarithms, $2{\cdot}7183 \cdots$, and

k = the relative deviate value of the variable

The binomial series

$$(p + q)^n = p^n + np^{n-1}q + \cdots + \frac{n!}{(n - r)!\,r!} p^{n-r}q^r + \cdots + q^n$$

Rule for deriving the numerical coefficient of a term from that of the preceding term:

> Multiply the numerical coefficient of the known term by the power of p in that term and divide by *one more than* the power of q in that term. The result is the numerical coefficient of the following term.

When n is large, the normal curve will give a good approximation to the binomial if np (or nq, whichever is smaller) is greater than 10.

The Poisson series

$$Ne^{-m}(1 + m + \frac{m^2}{2!} + \cdots + \frac{m^r}{r!} + \cdots)$$

This may be used as an approximation to the binomial when $m = np$ is less than 10, n being large and p being very small.

The Random Sampling Distributions of Statistics

The curves of frequency distribution for statistics calculated from samples of size N drawn at random from a specified supply have the following parameters:

Arithmetic mean, $\bar{x}$

$$\mu'_{\bar{x}} = \mu'_x$$

$$\sigma_{\bar{x}} = \frac{\sigma_x}{\sqrt{N}}$$

Curve is essentially normal except when N is too small to offset anormality in the supply.

Standard deviation, s_x

$$\mu'_{s_x} \longrightarrow \sigma_x$$

$$\sigma_s \longrightarrow \frac{\sigma_x}{\sqrt{2N}}$$

Curve approaches normality as N increases, essentially reaching that form when N is about 30 or more and the supply is normally distributed.

Correlation coefficient, r

$$\mu'_r \longrightarrow \rho$$

$$\sigma_r \longrightarrow \frac{1 - \rho^2}{\sqrt{N - 1}}$$

Curve approaches normality within only a small region of the (ρ,N) surface. For techniques circumventing this difficulty, see Chapter 11.

Proportions, p

$$\mu'_p = \pi$$

$$\sigma_p = \sqrt{\frac{\pi(1 - \pi)}{N}}$$

The distribution is that of the binomial $[\pi + (1 - \pi)]^N$, which is rather closely approximated by a normal curve if $N\pi$ is 10 or more.

Standard errors are estimates of the unknown true standard deviations of the random sampling distributions of statistics. They are commonly formed by replacing the unknown parameter with the "best estimate" of it, formulated from the sample. For large samples ($N > 30$) the following formulas are commonly employed:

$$\text{S.E.}_{\bar{x}} = \frac{s_x}{\sqrt{N}}$$

$$\text{S.E.}_{s_x} = \frac{s_x}{\sqrt{2N}}$$

$$\text{S.E.}_r = \frac{1 - r^2}{\sqrt{N}}$$

$$\text{S.E.}_p = \sqrt{\frac{p(1-p)}{N}}$$

Tests of Significance

(1) *The deviation of a mean*

$$k = \frac{\bar{x} - \mu_x'}{\sigma_{\bar{x}}}$$

$$(=) \frac{\bar{x} - \mu_x'}{\text{S.E.}_{\bar{x}}}$$

(Probability of k from Appendix I)

(2) *The difference between the means of two independent samples*

$$k = \frac{\bar{x} - \bar{y}}{\sqrt{\sigma_{\bar{x}}^2 + \sigma_{\bar{y}}^2}}$$

$$(=) \frac{\bar{x} - \bar{y}}{\sqrt{\text{S.E.}_{\bar{x}}^2 + \text{S.E.}_{\bar{y}}^2}}$$

(3) *Standard deviations and proportions.* Analogous tests follow the above procedures. Special techniques are required in small-sample work.

(4) *The deviation of r from zero*

(i) $N = 20$ or less

$$t = \frac{r\sqrt{N-2}}{\sqrt{1-r^2}}$$

(Probability of t from Appendix III)

(ii) $N > 20$

$z_r = \frac{1}{2}[\log_e(1+r) - \log_e(1-r)]$ (See Appendix II)

$k = z_r\sqrt{N-3}$

(Probability of k from Appendix I)

(5) *The difference between the correlation coefficients of two independent samples*

$$k = \frac{z_{r_1} - z_{r_2}}{\sqrt{N_1 + N_2 - 6}}$$

(Probability of k from Appendix I)

INDEX